Molecular Breeding **14:** 21–34, 2004.

Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags

Wencai Yang[1], Xiaodong Bai[2], Eileen Kabelka[1], Christina Eaton[1], Sophien Kamoun[3], Esther van der Knaap[1] and David Francis[1,*]

[1]Department of Horticulture and Crop Science; [2]Department of Entomology; [3]Department of Plant Pathology, The Ohio State University, Ohio Agricultural Research and Development Center, 1680 Madison Ave., Wooster, OH 44691, USA; [*]Author for correspondence (phone: 330-263-3893; fax: 330-263-3887; e-mail: francis.77@osu.edu)

Received 9 June 2003; accepted in revised form 19 November 2003

Key words: Fruit color, *Lycopersicon esculentum*, Mapping, QTL, Single nucleotide polymorphisms (SNP)

Abstract

Single nucleotide polymorphisms (SNPs) are useful for characterizing allelic variation, for genome-wide mapping, and as a tool for marker-assisted selection. Discovery of SNPs through *de novo* sequencing is inefficient within cultivated tomato (*Lycopersicon esculentum* Mill.) because the polymorphism rate is more than ten-fold 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 exploiting redundancy among EST data available for different varieties among 148,373 tomato ESTs, we have identified candidate SNPs for use within cultivated germplasm pools. 1,245 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. Sixty-six SNPs could be recognized by restriction enzymes, and subsequent experimental verification using restriction digestion or CEL I digestion confirmed 83% of the putative polymorphisms tested. SNPs between TA496 and Rio Grande have a high probability (53%) of detecting polymorphisms between other *L. esculentum* varieties. Twenty-six SNPs in 18 unigenes were mapped to specific chromosomes. Two SNPs, LEOH23 and LEOH37, were shown to be linked to quantitative trait loci contributing to fruit color within elite breeding populations. These results suggest that the growing databases of DNA sequence will yield information that facilitates improvement within the germplasm pools that have contributed to productive modern varieties.

Introduction

The use of wide crosses between cultivated varieties of tomato (*Lycopersicon esculentum* Mill.) and wild relatives (various *Lycopersicon* species) maximizes genetic variation and has lead to the discovery of new genes. However the emphasis on wide crosses has left a void in our ability to manipulate many traits of agricultural importance within elite breeding populations. A limitation to applying marker-assisted selection to the practice of breeding tomato varieties is that the low level of polymorphism between *L. esculentum* has precluded map coverage with sufficient density to fully use the power of modern biometrical techniques for trait discovery, genetic mapping and breeding. The lack of genetic markers that detect differences between elite breeding lines of tomato has prevented a detailed study of most traits of economic importance within genetic backgrounds that are relevant to plant breeders, growers, and processors.

There remains a need for molecular-marker systems that can exploit all polymorphisms.

Large-scale genome sequencing programs offer a potential solution to the scarcity of markers that can be used in elite populations. The tomato microsatellites or simple sequence repeats (SSRs) are an example of genetic markers that can be mined from existing sequence data (http://www.sgn.cornell.edu/). Single nucleotide polymorphisms (SNPs) are a second class of genetic markers that can be mined from sequence data and are useful for characterizing allelic variation, genome-wide mapping, and as a tool for marker-assisted selection. In the field of human genetics, SNPs are a major focus of efforts to increase the efficiency of mapping (International SNP Map Working Group 2001; Aerts et al. 2002; Balasubramanian et al. 2002; Chen et al. 2002) and are already being used for detection and mapping of a variety of diseases (Verhage et al. 2002; Sugimoto et al. 2002; Margiotti et al. 2002). In many crop plants, SNPs are present with sufficient frequency to offer an alternative for genetic mapping and marker-assisted selection. In maize, the frequency of polymorphisms in the US elite inbred lines is 1 SNP per 31 bp in non-coding regions, and 1 SNP per 124 bp in the coding regions (Ching et al. 2002). In soybean, a recent study of sequence diversity in 22 diverse genotypes found 1.64 SNPs per kb in coding regions, and 4.85 SNPs per kb in non-coding regions (Zhu et al. 2001). Kanazin et al. (2002) reported a rate of 1 mutation per 189 bases in barley. SNPs associated with traits have also been discovered in rice, soybean, and onion (Gupta et al. 2002). An advantage to using SNPs in plant breeding applications is that genotyping can be automated using single nucleotide primer extension assays (Giordano et al. 1999), thus offering a potential to increase both efficiency and throughput.

Although SNPs can be identified by sequencing selected DNA fragments, a practical limitation to this approach for tomato follows from the fact that the sequencing error rate is often higher than the polymorphism rate. The cost of SNP discovery through sequencing amplified fragments is therefore high even with reductions in the cost of sequencing. The objectives of the research described in this paper were to assess the potential of existing public databases for the discovery of polymorphisms. To date, the tomato genome project has resulted in a public database of 148,373 ESTs. Of these, 14.4% were derived from the variety Rio Grande or from Rio Grande × Moneymaker crosses (designated R11-12 and R11-13). Approximately 78.7% were derived from TA496, which has a processing tomato pedigree tracing to E6203. By comparing sequence data from Rio Grande and TA496 we assessed the potential to identify genetic differences between elite varieties. Polymorphisms discovered from this data mining were then applied to genetic studies within elite breeding populations.

Materials and methods

Identifying single nucleotide polymorphisms (SNPs)

Expressed sequence tags (ESTs) of *Lycopersicon esculentum* were obtained from the National Center for Biotechnology Information (NCBI) dbEST release 080902. The ESTs were downloaded in FASTA format as two distinct data sets based on the origin of varieties: TA496 and Rio Grande (including Rio Grande PtoR and the progeny of Rio Gande × Moneymaker, R11-12 and R11-13) using the *Entrez* search and retrieval system for nucleotide data and phrase searching (e.g., *Lycopersicon esculentum* [ORGN] AND EST AND TA496). FASTA formatted files were downloaded by directing the *cgi* text file to be saved on a local computer.

A set of scripts was written in Perl (version 5.6.0) to facilitate the manipulation and analysis of the FASTA sequence files. The EST entries extracted from the NCBI website were treated as input and modified by searching the description line for a specific string of “ESTxxxx” (where xxxx is a number), retaining only “ESTxxxx” as the entry name and adding a user-given extension name (TA496 or RioG) to the end of the entry names in the format of “ ESTxxxx.Extension”. Each EST sequence is therefore indexed to the NCBI database using ESTxxxx and to a variety based on the assigned Extension name.

ESTs of Rio Grande were assembled into a unique gene (unigene) contiguous sequence (contig) set using *Phrap* run on a workstation in the Linux operating environment. The *Phrap* output file was reduced to a file containing only contigs having 3 or more ESTs. These EST names were then re-integrated with the correct sequence data to form a file consisting of a contig number (Contigxxx) followed by three sequence data sets each with the “ESTxxxx.Extension” name to form a FASTA format sequence file. Next, a single sequence from each contig was chosen and searched against the EST database of TA496 using Basic Local Alignment Search Tool (BLAST).

Three EST sequences from the TA496 data set for each contig were selected using a program that takes the output file resulting from the BLAST search as the input. The top three hits from the BLAST output file were extracted and the information was stored in one file. The three TA496 sequences from the BLAST extractor output file were then combined with three EST sequences from the Rio Grande contig data set to create a data set with three EST sequences of Rio Grande (or related pedigrees) and three EST sequences of TA496. The resulting six EST sequences were aligned using the sequence alignment program ClustalX (1.8) to identify possible SNPs.

Confirmation of candidate SNPs

The SNPs detected by computer analysis were verified by PCR with restriction enzymes or by digestion with CEL I nuclease. Restriction enzyme cleavage sites at putative SNPs were detected in the sequences using Webcutter (Version 2.0, http://www.firstmarket-.com/cutter/cut2.html). Primers were designed using Primer 3 (Rozen and Skaletsky 2000) with the optimal PCR product length set between 150 and 600 bp.

SNP verification was based on the ability to detect expected polymorphisms in the DNA of E6203, TA496, Rio Grande, and Moneymaker. The confirmed polymorphisms were further screened for their potential in other crosses by testing a larger set of genotypes. DNA was isolated from 22 tomato varieties and breeding lines, a *L. esculentum* var *cerasiformae* plant introduction (PI) and 3 wild species *Lycopersicon* accessions (LA) using a modified CTAB isolation method as described previously (Kabelka et al. 2002). The varieties and wild accessions used were: E6203, TA496, Rio Grande, Moneymaker, NC84173, Fla7775, Fla7600, Ohio 9242, Ohio 8245, Ohio 7814, Ohio 88119, M 82, Sun 1642, Banana Legs, Sausage, Black Plum, Jersey Devil, San Marzano, Roma VF, Howard German, Hawaii 7998, Hawaii 7981, PI114490, LA1589, LA407, and LA716. PCR reactions were conducted in a 20μl reaction volume. Each reaction consisted of 10 mM Tris-HCl (pH 9.0 at room temperature), 50 mM KCl, 1.5 mM $MgCl_2$, 50 μM of each dNTP, 0.3 μM primers, 2 μl of 5 ng/μl genomic DNA template and 1 unit of *Taq* DNA ploymerase. Reactions were heated at 94 °C for 2 min followed by 36 cycles 1- min at 94 °C, 1-min at the suitable annealing temperature (Table 1), and a 2-min extension at 72 °C. Final reactions were extended at 72 °C for 5 min. Amplification was performed in a PTC-100™ programmable Thermal Controller (MJ Research, Inc. Watertown, MA). The PCR products detected as cut amplified polymorphic sequences (CAPS) were digested with specific restriction enzymes according to the manufacturer's protocol (Table 1). Fragments were separated using either 2% or 4% agarose gels (Amresco Biotechnology Grade 3:1 agarose, Solon, OH, USA), stained with ethidium bromide, and photographed using Syngene BioImaging Systems (Cambridge, UK).

Detection of polymorphisms using CEL I

The CEL I nuclease was partially purified using an approach modified from that described by Yang et al. (2000). Briefly, AEBSF replaced PMSF as a protease inhibitor in our protocol. Clean dry celery was homogenized at 4 °C using a professional series model JM211 juicer (Juiceman, Mt. Prospect, IL 60056, USA). One liter of juice was mixed with 30 ml of buffer A (0.1 M Tris-HCl pH 7.7, 100 μM AEBSF) and filtered through sterile cheesecloth three times in order to remove debris. All subsequent steps were performed at 4 °C with pre-chilled buffers, reagents and equipment as described by Yang et al (2000).

Fractions of crudely purified nuclease were assayed for CEL I activity using a heteroduplex template containing a loop of approximately 20 bp. Digestions were performed using each fraction from the crude preparation of CEL I and the heteroduplex template at 45 °C for 30 min. in 20 mM Tris-HCl pH 7.4, 25 mM KCl, and 10 mM $MgCl_2$ (Oleykowski et al. 1998). Digestion products were separated on 10% TBE-Urea polyacrylamide gels and stained with Sybr Gold (Molecular Probes, Eugene, OR, USA). Fractions containing the highest CEL I activity and minimal non-specific nuclease activity were retained.

A subset of SNPs (LEOH1, LEOH2, LEOH7, LEOH8, LEOH9, LEOH21, and LEOH22) were also confirmed using CEL I digestion of artificial heteroduplex templates (Table 1). In addition, primers amplifying loci TG23, TG91, TG47, TG134, TG236, TG242, TG246, TG359, TG609, CT59, CT93, CT118, CT167, CT168, CT182, and CT258 were tested for polymorphism using the CEL I assay. Heteroduplexes were formed by mixing equal amounts of amplified DNA from two tomato geneotypes, heating the DNA to 95 °C for 5 minutes to denature, and cooling to 53 °C to allow strands to re-anneal. This approach formed a mixture of homo and hetero-du-

Table 1. Summary of EST SNPs detected in *Lycopersicon esculentum*.

SNP	Chrom.	Rep. EST	Origin of EST	Codon substitution		Primer (5'----3')	Re. Enzyme	Temp.	Class
LEOH1.1	7	EST310638	TA496	non-synonymous	f:	TCC ACA TGA AGT AAT GGA CAC AG	NmuC I	60	verified
		EST253240	R11-12		r:	TTC TTC GTC AAG ATC GGG TA			
LEOH1.2	7	as above	as above			as above	CEL I		verified
LEOH2	unknown	EST478594	LA716		f:	CTT GAA GAT GGC CGA ACA CT	BsaW I	62	verified
		EST253241	R11-12		r:	CTG GTC TGG GGG AAT ACC TT	CEL I		verified
LEOH7	1	EST259392	R11-13		f:	TTC ATG TGC TGA CAT TCT TGC			
			LA1589		r:	TGA GTG TTG AGA CCC TTT GC	CEL I		verified
LEOH8.1	9	EST252213	TA496		f:	TCA AAT CAC AAA ATT AAC CTA TTC TTT	no enzyme		
		EST307657	Rio Grande		r:	GAC CAT TTT CCT AAC TCT TCA GG			
LEOH8.2	9	as above	as above		f:	CCA CTG ATC AAT GTG GTG GA	CEL I		verified
					r:	CAA CCA CAA ATG GCT CCT AAA			
LEOH8.3	9	as above	as above	synonymous		as above	CviJ I		not tested
LEOH8.4	9	as above	as above	synonymous		as above	HaeIII	55	verified
LEOH9.1	unknown	EST327297	TA496	non-synonymous	f:	GGC AAT GCC ACT GAC TTA CA	Aci I	55	verified
		EST287630	Rio Grande		r:	CTC TCT GCT GCT TCG GCT AC	CEL I		verified
LEOH9.2	unknown	as above	as above			as above	no enzyme		
LEOH10	4	EST547701	TA496	synonymous	f:	TGC CAG ATT GAC TGT GAA GG	BsaJ I	55	verified
		EST308333	Rio Grande		r:	GGA ACC CTG CAT TGT TCT TG			
LEOH11.1	unknown	EST328328	TA496		f:	TAT GTT GCT GCC CAG ACT CA	Hha I	55	not verified
		EST307804	Rio Grande		r:	ACA TCA TGA CCA ACC ATT CA			
LEOH11.2	unknown	as above	as above			as above	no enzyme		
LEOH12.1	unknown	EST549757	TA496		f:	CCA GAT GGG AGA TGG GTC TA	Bpi I	55	not verified
		EST308207	Rio Grande		r:	CAG CAG TAA CAC CAG GAG CA			
LEOH12.2	unknown	as above	as above			as above	Mse I	55	not verified
LEOH13.1	unknown	EST545360	TA496		f:	GGT AGA GTC CAA GCC CGA TT	no enzyme		
		EST287868	Rio Grande		r:	CGG ATC GAA TCC GTA GTC AC			
LEOH13.2	unknown	as above	as above	synonymous	f:	TGG CTG GTG ACA TTA TTG GA	Sty I	55	verified
					r:	CGG CAT CTT GCC ATG TAA TA			
LEOH14.1	unknown	EST542533	TA496		f:	GGG GTT GTT CAA CAT CTC ATT	no enzyme		
		EST283367	Rio Grande		r:	AAG TTG GCA TCC CAA GAC AT			
LEOH14.2	unknown	as above	as above	non-synonymous	f:	TCC GAG AGG CCA AGC TAT AA	TspR I	55	verified
					r:	GTA AGG ACG TTG TCC GAT CC			
LEOH15	2 & 3	EST475276	TA496	synonymous	f:	GCG GTT AAA CTC TCC CCA TC	Bpi I	55	verified
		EST285549	Rio Grande		r:	GTG TCC CAT CCG TAA TCA CC			
LEOH16.1	5	EST301659	TA496		f:	TGA ATT TTC TGT CAT CGT TGG	no enzyme		
		EST285093	Rio Grande		r:	TTT CGG AAT CTT TGT TGA ATT G			
LEOH16.2	5	as above	as above	non-synonymous	f:	TCG ACG CTG CAC AGA AAT AC	BsaW I	55	verified
					r:	TTC CTC CTC CTT ATC TCC TTC A			
LEOH16.3	5	as above	as above	non-synonymous		as above	BsaW I	55	verified
LEOH16.4	5	as above	as above	non-synonymous		as above	Hha I	55	verified
LEOH17.1	multiple	EST551464	TA496		f:	CAG ACG AGA AGC AAG TTG AGG	no enzyme		
		EST286054	Rio Grande		r:	CTA CCA CTG CGT GCT TTG AC			
LEOH17.2	multiple	as above	as above	non-synonymous		as above	Cac8 I	55	verified
LEOH17.3	multiple	as above	as above	synonymous		as above	BseN I	55	verified
LEOH17.4	multiple	as above	as above	non-synonymous		as above	Cac8 I	55	verified
LEOH17.5	multiple	as above	as above	non-synonymous		as above	Cac8 I	55	verified
LEOH19	12	EST353921	TA496	3' UTR	f:	AAG GCT CAG AAA GGG TCC AT	BsaB I	55	verified
		EST285582	Rio Grande		r:	TGA GTT CAT CAA CAC ATC ACA CA			
LEOH20.1	unknown	EST327354	TA496	non-synonymous	f:	CAG ACC TAA CAA GAC AGG CAA A	Hae III	55	verified
		EST284995	Rio Grande		r:	ATC AGG CAT GAC CAT GGA AG			
LEOH20.2	unknown	as above	as above			as above	no enzyme		
LEOH21	unknown	EST298805	TA496		f:	ACT CCA CCT GTT GCC AAG AC	no enzyme		
		EST308753	Rio Grande		r:	CCA ACA AGC ATC AAG TCA CC	CEL I		verified
LEOH22	unknown	EST511660	TA496		f:	TCG AGA GTT GCT GCT GAA TTT	no enzyme		
		EST286802	Rio Grande		r:	AAT GTG CCT TTT TGC AAT GAT	CEL I		verified

Table 1. Continued.

SNP	Chrom.	Rep. EST	Origin of EST	Codon substitution		Primer (5'----3')	Re. Enzyme	Temp.	Class
LEOH23.1	2	EST546919	TA496	5' UTR	f:	GAG AGA AAA AGG GCA CAA GG	Msp I	56	verified
		EST256088	R11-12		r:	ACC GAC AAA CGC ATA GAT CA			
LEOH23.2	2	as above	as above		f:	CTA TGC GTT TGT CGG TCG T	BspM I		not tested
					r:	CAA GGT AGT TGA AGG TAT GAC CA			
LEOH23.3	2	as above	as above	synonymous	f:	as above	Tsp509 I	54	verified
LEOH24	unknown	EST587500	TA496		f:	CTG GTG AAT ATG GCG GTC TT	Mse I		not verified
		EST281057	Rio Grande		r:	TCT CGT GAA GTG GCA TCA AG			
LEOH25.1	9	EST511738	TA496	synonymous	f:	GGA GGA AAT AGG GTT TCT AGG G	Hinc II	56	verified
		EST285647	Rio Grande		r:	AAT GGC CTG GCT AAT CTG TG			
LEOH25.2	9	as above	as above			as above	Bbv I		not tested
LEOH25.3	9	as above	as above	non-synonymous		as above	BasA I	56	verified
LEOH26	unknown	EST329684	TA496	non-synonymous	f:	GAA GAT TCG GAG GTC AAA CG	Fok I	56	verified
		EST279770	Rio Grande		r:	AAC TCC TCA ACT GCC TCA GC			
LEOH27.1	unknown	EST584485	TA496		f:	CCA AAT TTC CAT CTC CCA TT	Mse I	56	not verified
		EST261932	R11-13		r:	ATG GCC CTT CCT TTG TTT CT			
LEOH27.2	unknown	as above	as above		f:	GGG AAG CAT AAG TGC AGC TC	no enzyme		
					r:	TCC CCC ATA ATT TCT TAT CGT			
LEOH28	unknown	EST273353	TA496		f:	CGC GGA GTT CTG TTA GCT TC	Taq I	56	not verified
		EST253877	R11-12		r:	GCC GAC GAA TTA CGA ACA TC			
LEOH29.1	unknown	EST243853	TA496	non-synonymous	f:	CCT CCA TGA CCG ATG CTA CT	Alu I	56	verified
		EST255173	R11-12		r:	TAG TGA TTC CTC CGT GGA CA			
LEOH29.2	unknown	as above	as above			as above	Mae II		not tested
LEOH30.1	unknown	EST399721	TA496		f:	CAG GTT TCA GCT ACT GGA TTT TG	Taq I	53	not verified
		EST285542	Rio Grande		r:	TCT ACA TGG ACC ACA CCA TGA			
LEOH30.2	unknown	as above	as above			as above	no enzyme		
LEOH31.1	9	EST583372	TA496	synonymous	f:	TTG CAA TGG CTT CTC TCC TC	Tsp509 I		not tested
		EST308897	Rio Grande		r:	ACT TGT CCG TTT CTC GCT TG			
LEOH31.2	9	as above	as above	synonymous		as above	Tsp509 I		not tested
LEOH31.3	9	as above	as above	synonymous		as above	Msp I	51	verified
LEOH31.4	9	as above	as above	non-synonymous		as above	Mse I	51	verified
LEOH31.5	9	as above	as above		f:	TGT TGA TGT CTG GTC CAT TTC T	no enzyme		
					r:	CCC TGC CCA AAC ATC TAA AA			
LEOH31.6	9	as above	as above	synonymous		as above	Taq I	54	verified
LEOH31.7	9	as above	as above	synonymous		as above	Mse I	54	verified
LEOH31.8	9	as above	as above			as above	no enzyme		
LEOH31.9	9	as above	as above	non-synonymous		as above	Aci I	54	verified
LEOH31.10	9	as above	as above	non-synonymous		as above	Alu I	54	verified
LEOH32.1	9	EST358606	TA496	synonymous	f:	TGG TGT GGA TCC TGC TGT TA	Hae III	56	verified
		EST256921	R11-12		r:	TGG AAA TCA CAC CAA AAC GA			
LEOH32.2	9	as above	as above	synonymous		as above	Dra III	56	verified
LEOH33.1	9	EST471439	TA496		f:	GAG TGT GAA GGG AAG GCA CT	BesN I		not tested
		EST262714	R11-13		r:	TTT GGA ATC GGA AGA ACC AG			
LEOH33.2	9	EST471439	TA496	non-synonymous	f:	TGA GGA AGC TTG CTG ACA AA	Mse I	54	verified
		EST262714	R11-13		r:	GCC TTT ATC TTT TAA AGC TGC AAT			
LEOH33.3	9	as above	as above	non-synonymous		as above	Tsp509 I	54	verified
LEOH33.4	9	as above	as above	non-synonymous		as above	Cac8 I	54	verified
LEOH33.5	9	EST471439	TA496		f:	GTT TTT GCT GGG GTG GAA AG	no enzyme		
		EST262714	R11-13		r:	CTC AGG AGG AGT TGG ACG AT			
LEOH34.1	9	EST435427	TA496		f:	CGT TTC ACC AAT TGC ACT GA	no enzyme		
		EST260581	R11-13		r:	CTT ATG TAT CGC GGG CCT TC			
LEOH34.2	9	as above	as above	synonymous		as above	Tsp509 I	56	verified
LEOH34.3	9	as above	as above			as above	no enzyme		
LEOH34.4	9	as above	as above			as above	no enzyme		
LEOH34.5	9	as above	as above			as above	no enzyme		
LEOH34.6	9	as above	as above	synonymous		as above	NmuC I	56	verified

Table 1. Continued.

SNP	Chrom.	Rep. EST	Origin of EST	Codon substitution		Primer (5'----3')	Re. Enzyme	Temp.	Class
LEOH34.7	9	as above	as above			as above	Fok I	56	not verified
LEOH35.1	9	EST549543	TA496		f:	CAT CAG CCT CGC TCT CTT CT	Ema1104 I		not tested
		EST280079	Rio Grande		r:	CAA ACT GCA AGC CAT TTG AA			
LEOH35.2	9	as above	as above			as above	no enzyme		
LEOH35.3	9	as above	as above	5' UTR		as above	BseN I	56	verified
LEOH35.4	9	as above	as above			as above	no enzyme		
LEOH35.5	9	as above	as above	non-synonymous		as above	Taq I	56	verified
LEOH35.6	9	as above	as above			as above	no enzyme		
LEOH35.7	9	as above	as above	synonymous		as above	BsrB I	56	verified
LEOH35.8	9	as above	as above	synonymous		as above	Cac8 I	56	verified
LEOH36	1	EST550593	TA496	non-synonymous	f:	TCA CAA AAA TGG CGA TGA GA	Bcl I	56	verified
		EST279973	Rio Grande		r:	CCA CCT GTG GAT CCT TGA CT			
LEOH37	4	EST319984	TA496	3' UTR	f:	TTG ATA TAT TCC ATG TGT GTC TC	NmuC I	51	verified
		EST258553	R11-13		r:	AAC TAC AAA TTA ACA AAC TTA AAT GG			
LEOH38.1	unknown	EST283101	TA496		f:	TGG GAA GAT TAT GCA TGC TG	no enzyme		
		EST253060	R11-12		r:	GCC CTT CTG AAT TTT CGA GTC			
LEOH38.2	unknown	EST283101	TA496	non-synonymous	f:	CAA GGT TGT GGC TAT GCT CA	Aci I	56	verified
		EST253060	R11-12		r:	ACC TCA GCA GGA TTG ACG AG			
LEOH38.3	unknown	as above	as above			as above	no enzyme		
LEOH39	unknown	EST552218	TA496		f:	AGA GAG TGG TGC AAG TTA G	EcoR II	56	not verified
		EST280326	Rio Grande		r:	CAT AGG CAC AGT AAT GAG AT			
LEOH40.1	7	EST244875	TA496	non-synonymous	f:	TGA GTT GGT GAA CCA TGG AA	NmuC I	56	verified
		EST253240	R11-12		r:	CCA AAG TTG GGA CCT TTT GA			
LEOH40.2	7	as above	as above			as above	no enzyme		
LEOH41	unknown	EST554628	TA496		f:	GAA ACA GCT GGG AAT TTT GC	no enzyme		
		EST263182	R11-13		r:	CCG TTG TGT TTG CTA TGT TCA			
LEOH42.1	unknown	EST527482	TA496		f:	GGC GGA TGT TCA GAG AAG AG	BseR I		not tested
		EST308868	Rio Grande		r:	TGG ACC TCA TCT TTG GGT TC			
LEOH42.2	unknown	as above	as above			as above	BseR I		not tested
LEOH43	unknown	EST328521	TA496		f:	TTG CTC CCG AGA GTC TTG AA	Ema1104 I		not tested
		EST289791	Rio Grande		r:	TTA CCA AAG CAA TGC CAC CT			
LEOH44	unknown	EST467087	TA496		f:	GTG CAT TCA CGA ATT CCA CA	SfaN I		not tested
		EST261539	R11-13		r:	TTA GAA CCT CCC CCA AAT CC			
LEOH45	unknown	EST526153	TA496		f:	GCA AAC GGA GTT TCT TCG AG	no enzyme		
		EST260297	R11-13		r:	TGC ACT TTT CTT GCT CCT GA			
LEOH46	unknown	EST532932	TA496		f:	GAC GTG CTA AAA GGG ACT CG	Bsg I		not tested
		EST260249	R11-13		r:	TTC ATG AAG CAA CGA TCC AA			
LEOH47.1	unknown	EST542104	TA496		f:	TGT TCA CAC AGA ATA AGT TGC TC	Hph I		not tested
		EST254278	R11-12		r:	AAT TGC CCA TTT CAA AGC TG			
LEOH47.2	unknown	as above	as above			as above	no enzyme		
LEOH48	unknown	EST412384	TA496		f:	ACC GGA AAT TCA GTT CAT GC	no enzyme		
		EST284722	Rio Grande		r:	AGC ACG AAC ACC AAG ACC AT			
LEOH49	unknown	EST243322	TA496		f:	CCA CCT TCA AAC AAG TCA GC	no enzyme		
		EST257061	R11-12		r:	GCG CCT GCC ATA TAT TTT GA			
LEOH50.1	unknown	EST326794	TA496		f:	GGT GCC AGA TTC AGA TGT CA	no enzyme		
		EST253341	R11-12		r:	CTC TGT TCC ATC ACC GGA GT			
LEOH50.2	unknown	as above	as above			as above	no enzyme		
LEOH51	unknown	EST533262	TA496		f:	AAG GGC TGG TGT GAA AGC TA	no enzyme		
		EST261807	R11-13		r:	CAT TTC CAA AAA CTC CAG CA			

plex DNA that served as a template for CEL I nuclease digestion. Digestion products were separated on 10% TBE-Urea polyacrylamide gels and stained with Sybr Gold for visualization.

Non-synonymous and synonymous substitution

Non-synonymous and synonymous mutations were detected within ESTs containing SNPs by scanning sequences for open reading frames using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Putative ORFs were used to search the NCBI data for homology between amino acid sequences (BLAST-P). Open reading frames showing a high percentage match to known genes were assumed to be correct, and the amino acid sequences within each contig were then aligned using ClustalX (1.8) to determine whether substitutions were synonymous or non-synonymous.

Genetic mapping of SNPs

Two populations were used to map the SNPs. The first population (population 1) was a set of *L. pennellii* LA716 introgression lines (ILs). Each line is homozygous for a single chromosome segment derived from LA716 and delineated by RFLP markers introgressed from *L. pennellii* into *L. esculentum* cultivar M82, such that the entire wild species genome is represented in a group of 50 lines (Eshed and Zamir 1995). The second population was an F_2 population (Population 2) consisting of 46 individuals derived from a cross of LA1589 (*L. pimpinellifolium*) and Sun1642 (*L. esculentum*). The SNP markers were combined with RFLP markers placed on the same population (van der Knaap and Tanksley 2001) to construct a linkage map using the Kosambi mapping function of Mapmaker (Lander et al. 1987).

Color measurement

Two populations derived from *L. esculentum*×*L. esculentum* crosses were analyzed for the association of SNPs and loci that affect fruit color. The first population was derived from Ohio 8245 and Ohio 2349 (Kabelka 2001) and consisted of 160 F_2 individuals. The second population consisted of 80 F_2 individuals derived from crossing Ohio 1023 and Ohio 7814. Populations were grown in the field using conventional practices (Precheur 2000), and twenty-four fruit were harvested from each plant for objective measurement of color as described by Sacks and Francis (2001).

Numeric 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). The L* coordinate indicates darkness or lightness of color and ranges from black (0) to white (100). Coordinates, a* and b*, indicate color directions: +a* is the red direction, −a* is the green direction, +b* is the yellow direction and −b* is the blue direction. Chroma (saturation or vividness of color) and hue (the basic tint of color) are derived from a* and b*. Chroma is calculated as $(a^{*2} + b^{*2})^{1/2}$. As chromaticity increases, a color becomes more intense; as it decreases a color becomes duller. A minimum color CIELAB difference of 1 unit is perceptible to a human observer depending on the L* value, background color, and lighting (Berger-Schunn 1994). Hue is an angular measurement, calculated as $(180/\pi)[\cos^{-1}$ (a*/chroma)] for positive values of b*, and is defined as starting at the red +a* axis at 0 degrees. A hue angle of 45 degrees would be orange-red in color, whereas 90 degrees would be yellow. Perception of hue angle differences will depend on the chroma with differences more detectable at higher chroma. In general, and based on the assumption that there are approximately 160 distinguishable hues, a hue angle difference of 2.5 is detectable (Hardin 1990).

Genotyping and statistical analysis for marker-trait association

All SNPs were examined for polymorphism against parents of elite breeding populations. A total of nine PCR-based markers (based on TG and CT sequences) including three newly identified SNPs were tested in the OH1023×OH7814 population. Sixteen PCR-based markers including five newly identified SNPs were tested in the OH8245×OH2349 population. Genotyping was performed as described above for SNP verification.

Statistical analyses were performed using the GLM procedure of SAS (Statistical Analysis System version 8.1, SAS Institute, Cary, NC). The statistical models and the rationale for these models have been described in detail previously (Sacks and Francis 2001; Kabelka et al. 2002). Linkage relationships between the genotypic classes of each molecular marker with hue, L, and chroma within populations were determined with molecular marker considered as a fixed

Table 2. Distribution of substitution types among confirmed SNPs.

Substitution type	Base substitution	No. of occurrence	Percentage
Transition	A/G	5	11.6
	G/A	9	20.9
	C/T	8	18.6
	T/C	3	7.0
Sub-total		25	58.1
Transvertion	A/T	1	2.3
	A/C	6	14.0
	C/G	3	7.0
	G/C	1	2.3
	G/T	1	2.3
	T/G	1	2.3
	T/A	4	9.3
Sub-total		17	39.5
Insertion/ deletion	G	1	2.3
Total	43	100	

effect whereas replicated measurements of fruit color and genotypes were considered as random effects. The statistical model tested accounted for variation within fruit and within F_2 plant, degrees of freedom were calculated via the Satterthwaite approximation, and the genotype within marker variation was specified as the error term for the F-statistic. The marker-trait analysis was therefore more conservative than statistical approaches that rely only on the mean fruit color for each F_2 plant (Sacks and Francis 2001; Kabelka et al. 2002). Significant ($p < 0.05$) differences in marker class means were interpreted as evidence for linkage of a marker to a locus controlling hue, L, or chroma. Because our model for marker-trait analysis accounts for within fruit and within plant variation, total phenotypic variation explained by each marker was calculated by partitioning variance components using the VARCOMP procedure of SAS and restricted maximum likelihood (REML).

Results

SNPs between TA496 and Rio Grande

A total of 138,093 EST sequences derived from different tissues of either TA496 or Rio Grande including the progeny of a cross between Rio Grande and Moneymaker (R11-12 and R11-13) were obtained from the NCBI. 21,382 (14.4% of the total and 15.5% of the downloaded sequences) ESTs of Rio Grande, R11-12 and R11-13 were assembled into 2,635 unigenes. Applying a cut-off of at least three sequences for each contig resulted in a data-set consisting of 1,504 contigs that provided the basis for further analysis. The automated and random selection of a single sequence from each contig for BLAST against the EST database of TA496 identified 1,245 contigs with three or more sequences common to the two sets of sequence data. The 138,093 ESTs were therefore reduced to 1,245 contigs for use in identifying potential SNPs. Among these sequences, forty-four unigenes showed 101 potential polymorphisms, two of which were putative insertion/deletions (indel) mutations.

Sixty-six candidate SNPs could be recognized by available restriction enzymes. Fifty-two SNPs in 33 unigenes were selected for PCR and restriction digestion analysis. For initial verification four varieties, TA496, Rio Grande, E6203, and Moneymaker, were used for PCR and the products were digested by the appropriate restriction enzyme. Forty-three (82.7%) candidate SNPs (including 1 indel) in 24 unigenes were confirmed (Table 1). The distribution of genetic changes was: 58.1% transitions, 39.5% transversions and 2.3% indels (Table 2). Seventeen out of 23 SNPs were confirmed using CEL I digestion (Figure 1). Using a subset of SNPs verified in this work and

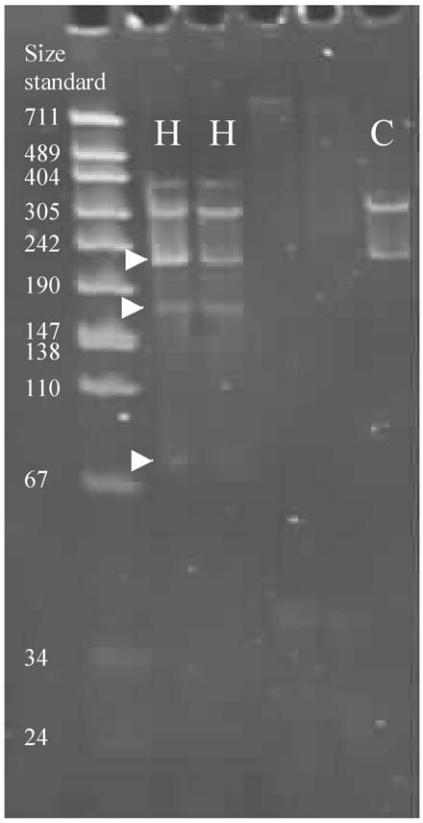

Figure 1. Example of SNP detection with CEL I digestion of PCR-amplified DNA of LA1589 and 86120 (*L. esculentum*). C: control consisting of undigested heteroduplex; H: Heteroduplex DNA treated with CEL I. The heteroduplex DNA template is formed from denaturation and renaturation of two distinct genotypes and consists of a mixture of homoduplex DNA and heteroduplex DNA. CEL I treatment results in digestion of only the heteroduplex portion of the template leaving homoduplex DNA undigested (top arrow). Digestion products of expected size are indicated by the lower two arrows.

known SNPs in TG and CT sequences as a basis of comparison, we estimate that the CEL I assay detected 74% of true SNPs under the conditions employed (data not shown).

The frequency of polymorphisms between TA496 and Rio Grande is 1 SNP in approximately 8,500 bases. We detected an average of 1 SNP per 15 genes (based on an estimate of 83 confirmed SNPs per 1,245 unique genes). However, the average number of SNPs per polymorphic EST was 1.79; 43.2 % of polymorphic ESTs had only one SNP, 34.1 % had two, 6.8% had three, 4.5% had four, and five ESTs (11.4%) had 5 or more SNPs. Given the distribution of SNPs per gene, an estimation of the number of polymorphisms between TA496 and Rio Grande is 1 SNP for every 28 genes.

Of the 43 SNPs confirmed by restriction digestion, 23 are non-synonymous substitutions and 16 are synonymous substitutions. The remaining 4 SNPs appeared to be in regions of the EST sequence that are not translated (Table 1).

The presence of confirmed SNPs in other tomato germplasm

To test if the SNPs identified between TA496 and Rio Grande are also polymorphic among other *L. esculentum* varieties, an additional 19 varieties representing fresh market varieties, processing varieties, heirloom varieties, and breeding lines were compared. Of the 43 SNPs between TA496 and Rio Grande that were confirmed with restriction digest, 23 also showed polymorphisms among other *L. esculentum* varieties (data not show). This indicated that the SNPs discovered between Rio Grande and TA496 had a high probability (53.5%) of detecting SNPs between other *L. esculentum* varieties.

SNPs identified in *L. esculentum* were also observed in the three wild species: LA716 (*L. pennellii*), LA407 (*L. hirsutum*), and LA1589 (*L. pimpinellifolium*). SNPs present between TA496 and Rio Grande had 82.5% polymorphism rate with LA716, 80% with LA407, and 67.5% with LA1589. Occasionally polymorphisms were detected between *L. esculentum* and the wild species that were not detected based on the computer analysis. For example, an indel polymorphism was detected with LEOH7 between *L. esculentum* and LA1589 and a SNP was detected with LEOH2 between *L. esculentum* and LA716 (Table 1).

Map position of SNPs

Of the 43 confirmed SNPs, 12 showed specific polymorphisms between LA716 and M82, 4 showed specific polymorphisms between LA1589 and Sun1642, and 10 showed polymorphisms in both populations. These 26 SNPs belonged to 18 unigenes, and 16 were mapped to a specific chromosome (Table 1, Figure 2). Map positions of most SNPs in common to both populations were consistent, e.g., LEOH37 was mapped to IL4-3 using population 1 and mapped to chromosome 4 using the population 2. LEOH16 mapped between markers CT93 and TG96 on chromosome 5 using population 2 but none of the IL lines

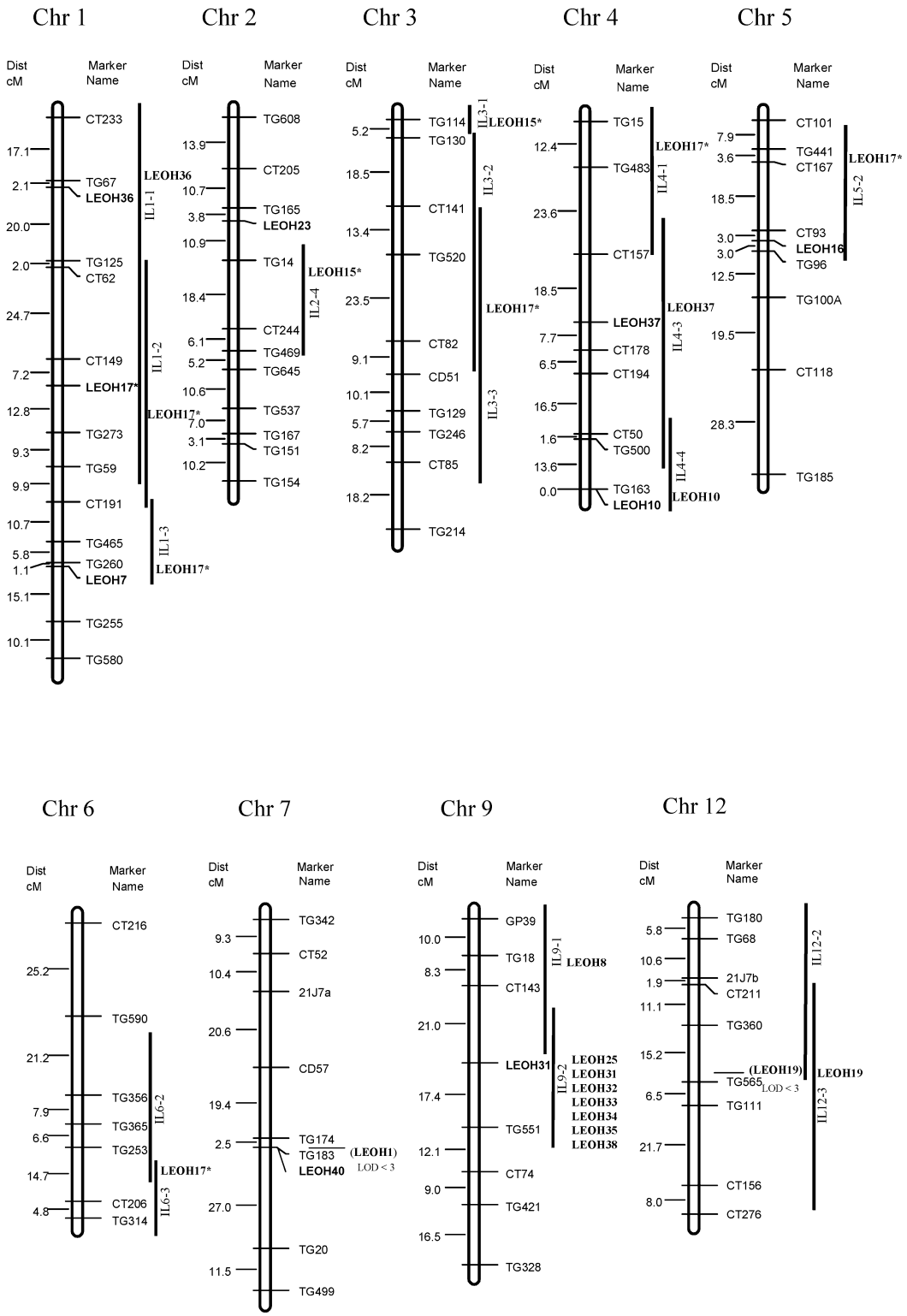

Figure 2. Map position of SNPs. SNPs mapped in the F_2 population derived from a cross of LA1589 (*L. pimpinellifolium*) and Sun1642 (*L. esculentum*) are indicated relative to framework markers on chromosomes. Bold lines to the right of each chromosome indicate the positions of *L. pennellii* LA716 introgression lines (ILs). SNP markers only 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.

Table 3. SNPs associated with lightness-darkness of color (L) and intensity of color (Chroma) in two elite breeding populations.

			L			Chroma		
Marker	Population	Genotypic Class	Mean	p	Vp	Mean	p	Vp
LEOH23	OH1023×OH7814			0.022	0.146		ns	ns
		OH1023	42.87			36.58		
		OH7814	40.72			37.32		
		Heteroz.	41.78			36.93		
LEOH37	OH8245×OH2349			ns	ns		<0.0001	0.216
		OH8245	40.16			39.04		
		OH2349	40.11			36.49		
		Heteroz.	39.77			38.37		

Significance (p) of single marker-trait analysis is based on an F-test using Gen(marker) variation as the error term; proportion of total phenotypic variance explained is indicated by Vp; ns=not significant at 0.05.

in population 1 showed the polymorphism that was detected between LA716 and M82. Likewise, LEOH40 mapped on chromosome 7 in population 2, but the polymorphism detected between parents was not found in the segregating IL population. Two unigenes, LEOH15 and LEOH17, detected multiple gene families and could not be mapped to a specific chromosome. LEOH15 amplified a *CAB* gene with family members that were mapped to chromosomes 2 and 3, consistent with the location of *CAB*1 and *CAB*3 respectively. LEOH17 amplified an *Adh* gene that was mapped to 5 chromosomes in the IL population and only chromosome 1 using population 2. Map positions were consistent with the location of *Adh*1 and *Adh*2 (Tanksley and Jones 1981). Although the map positions of a subset of the *CAB* and *Adh* genes were consistent with reference maps, it is also possible that the few inconsistent results between the two mapping populations are due to IL lines containing small introgressions from other chromosomes and or small gaps from the introgressed *L. pennellii* genome (Bonnema, et al., 2002). The mapped SNPs cover 9 of 12 tomato chromosomes, with half of them placed on chromosome 9.

Identifying SNPs associated with fruit color

Single marker-trait analysis of the F_2 populations involving elite breeding lines revealed that polymorphisms detected by LEOH37 and LEOH23 were significantly associated with loci that affect components of tomato fruit color (Table 3). F_2 plants with the Ohio 8245 allele marked by LEOH37 showed an increase in chroma that corresponds to twice the level detectable by an average observer. This locus explained 21.3% of the total phenotypic variation for chroma and probably corresponds to the locus on chromosome 4 described based on a RAPD polymorphism detected by OPBB-09 (Kabelka 2001). F_2 plants with the Ohio 7814 allele of LEOH23 showed decreasing L values and explained 14.6% of the total phenotypic variation for L. Again, this change in L corresponds to two-fold the difference perceptible to a human observer (Berger-Schunn, 1994; Hardin, 1990). Thus, the SNPs were useful for detecting two QTL for color and may have applications for marker-assisted selection within populations derived from elite *L. esculentum* varieties.

Discussion

Large-scale sequencing of Expressed Sequence Tags and complete genomes offers information of use to plant breeding programs. With the completion of the first crop genome sequencing projects (Goff et al. 2002; Yu et al. 2002) the potential for plant breeding to be impacted by new technology has never been greater. In tomato, sequencing projects offer a potential solution to the scarcity of markers that can be used in elite breeding populations. Of special interest is the ability to discover DNA polymorphisms by mining sequence data (Smulders et al. 1997; Bredemeijer et al. 2002).

The frequency of single nucleotide polymorphisms that we detected is considerably lower than reported for maize, wheat, barley, and soybean. Not surprisingly it is also lower than the one SNP per approximately 100 bases that was detected between *L. pennellii* and *L. esculentum* (Suliman-Pollatschek et

al. 2002). However, this result must be interpreted with care as TA496 and Rio Grande are both determinate "roma" style tomatoes and therefore do not fully represent the diversity within cultivated germplasm. Based on SSR markers, TA496 and Rio Grande represent less than 37% of the genetic variation in cultivated tomato (unpublished data). In contrast, soybean and maize studies examined SNPs in germplasm from more diverse populations (Zhu et al. 2001, Ching et al. 2002). Based on the frequency of SNPs that we detected and considering the estimate of 35,000 genes in tomato (Van der Hoeven et al. 2002), we may expect as many as 2,300 polymorphisms between genes of these two *L. esculentum* varieties. Given the average of 1.79 SNP per gene, we expect as many as 1,284 unique genes could be polymorphic between TA496 and Rio Grande.

One limitation to detecting SNPs is the need to sequence alleles from both parents. With the occurrence of SNPs falling below the sequence error rate, this approach is potentially costly. A second approach is to use the CEL I assay to detect SNPs. This approach will miss as many as 25% of SNPs, but may offer a high-throughput option. Preliminary results using CEL I suggest that SNP detection in non-coding DNA will be considerably more efficient than SNP detection in coding regions(data not shown). Thus, it is not unreasonable to assume that marker coverage based on SNPs and appropriate for interval mapping could be achieved for crosses within cultivated germplasm.

Several factors that affected the success of "*in silico*" polymorphism detection could be addressed with further analysis or experimentation. First, the sequencing error was approximately 0.2% in our data set which excluded the extreme 5' and 3' portions of sequence runs. Thus the sequence error was roughly 17 fold higher than the true polymorphism rate. Public access to EST sequence trace files or *Phred* quality scores may allow for more efficient SNP discovery by permitting the use of quality information as a substitute for sequence redundancy. Second, the current EST data set is heavily skewed towards TA496 sequences thus restricting the data set available for comparisons to 15.5% of sequence. Sequencing efforts aimed at obtaining a more balanced data set based on variety of origin will permit more effective discovery of polymorphisms. Finally, although we relied on three-fold redundancy there were still 17% of candidate SNPs that could not be confirmed. We believe that either shared sequencing error, closely related multi-gene families, intron disruption of restriction sites in the EST, or variety source differences contributed to the detection of SNPs "*in silico*" that could not be confirmed. The successful identification of markers that are polymorphic within cultivated germplasm and the potential for many more suggests that continued mining of sequence data for SNPs will be productive.

It is possible to interpret the sequence and mapping data in light of the role that selection may play in maintaining SNPs within the germplasm pool of cultivated tomato. Although the mapping of SNPs identified markers on nine of the twelve tomato chromosomes, many SNPs mapped to chromosome 9. The clustered distribution may reflect selection pressure on chromosome 9 that differentiates TA496 and Rio Grande. It is entirely possible that the introgression of Tm-2 on chromosome 9 of TA496 carried a linked block of genes from the wild species donor of resistance. In support of this hypothesis, only one of the eight genes that map to chromosome 9 was polymorphic between other *L. esculentum* varieties (excluding Fla 7775 which also contains Tm-2). Of the remaining ten markers that were mapped, dispersed regions of the genome were covered and nine were polymorphic in other *L. esculentum* varieties. These results suggest that further SNP identification will not only tag introgressions, but also provide distribution across the genome.

The utility of the *L. esculentum* SNPs for breeding and genetic applications is validated by the demonstration that LEOH23 and LEOH37 are associated with QTL contributing to fruit color within breeding populations of tomato. We had previously shown that genetic variation for color exists within such elite breeding populations and that this variation could not be explained based on known genes (Sacks and Francis, 2001). The amino acid sequence and map position on chromosome 4 demonstrate that LEOH37 is LeMT3, a member of the type II metallothionein gene family in tomato (Giritch et al. 1998). A QTL linked to LeMT3 explained 21.6% of the total phenotypic variation for chroma, the intensity of color, and does not correspond to previously described genes known to affect color in tomato. A second SNP, LEOH23, is associated with a locus that affects 14.6% of the phenotypic variation for the lightness to darkness of tomato fruit. LEOH23 maps to chromosome 2, the same chromosome that contains PSY2 (Bartley and Scolnik, 1993) and PHYE (van Tuinen et al, 1997). The function of PSY2 and PHYE are sufficient to consider these genes as "candidate loci" for QTL that

contribute to fruit color. However, the transcripts of the PSY2 phytoene desaturase are more abundant in mature leaves than fruit (Bartley and Scolnik, 1993), and this gene has not been considered an important contributor to carotenoids in the fruit. Most importantly, chromosome 2 has not previously been associated with loci that influence color and has not been actively targeted by plant breeders seeking to improve fruit quality. The SNP markers identified in this study will therefore be useful in marker-assisted selection for color.

Acknowledgement

The authors thank Dr. Guoliang Wang (Department of Plant Pathology) for providing access to his workstation and to the Molecular and Cellular Imaging Center for assistance with bioinformatics. Salaries and research support were provided by state and federal funds appropriated to The Ohio State University, Ohio Agricultural Research and Development Center, and grant funds from the Mid-American Food Processors. The mention of firm names or trade products does not imply that they are endorsed or recommended by The Ohio State University over other firms or similar products not mentioned.

References

Aerts J., Wetzels Y., Cohen N. and Aerssens J. 2002. Data mining of public SNP databases for the selection of intragenic SNPs. Human Mutation 20: 162–173.

Balasubramanian S., Harrison P., Hegyi H., Bertone P., Luscombe N., Echols N., McGarvey P., Zhang Z. and Gerstein M. 2002. SNPs on human chromosomes 21 and 22 – analysis in terms of protein features and pseudogenes Pharmacogenomics 3(3): 393–402.

Bartley G.E. and Scolnik P.A. 1993. cDNA cloning, expression during development, and genome mapping of PSY2, a second tomato gene encoding phytoene synthase. J. Biol. Chem. 268(34): 25718–25721.

Berger-Schunn A. 1994. Practical color measurement: A primer for the beginner, a reminder for the expert. Wiley, New York, New York, USA.

Bonnema G., Berg P. and Lindhout P. 2002. AFLPs mark different genomic regions compared with RFLPs: a case study in tomato. Genome 45: 217–221.

Bredemeijer G.M.M., Cooke R.J., Ganal M.W., Peeters R., Isaac P., Noordijk Y., Rendell S., Jackson J., Röder M.S., Wendehake K., Dijcks M., Amelaine M., Wickaert V., Bertrand L. and Vosman B. 2002. Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theor. Appl. Genet. 105: 1019–1026.

Chen L.Y.Y., Lu S.H., Shih E.S.C. and Hwang M.J. 2002. Single nucleotide polymorphism mapping using genome-wide unique sequences. Genome Research 12: 1106–1111.

Ching A., Caldwell K.S., Jung M., Dolan M., Smith O.S., Tingey S., Morgante M. and Rafalski A.J. 2002. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3(1):19.

Eshed Y. and Zamir D. 1995. An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. Genetics 141: 1147–1162.

Giordano M., Oefner P.J., Underhill P.A., Cavalli Sforza L.L., Tosi R. and Momigliano Richiardi P. 1999. Identification by denaturing high-performance liquid chromatography of numerous polymorphisms in a candidate region for multiple sclerosis susceptibility. Genomics 56: 247–253.

Giritch A., Ganal M., Stephan U.W. and Baumlein H. 1998. Structure, expression and chromosomal localisation of the metallothionein-like gene family of tomato. Plant Molecular Biology 37: 701–714.

Goff S.A., Ricke D., Lan T.H. et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296: 92–100.

Gupta P.K., Roy J.K. and Prasad M. 2002. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Current Science 80: 524–535.

Hardin C.L. 1990. Why color? Perceiving, Measuring, and Using Color. Soc. Photo-Optical Instrumentation Engineers, 1250: 293–300.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933.

Kabelka E. 2001. Discovery and introgression of beneficial loci from *Lycopersicon hirsutum*, LA407, a wild species of tomato. PhD dissertation, Horticulture and Crop Science, The Ohio State University, Ohio, USA.

Kabelka E., Franchino B. and Francis D.M. 2002. Two loci from *Lycopercosin hirstum* LA407 confer resistance to strains of *Clavibacter michiganensis* subsp *michiganensis*. Phytopathology 92: 504–510.

Kanazin V., Talbert H., See D., DeCamp P., Nevo E. and Blake T. 2002. Discovery and assay of single-nucleotide polymorphisms in barley (Hordeum vulgare). Plant Molecular Biology 48: 529–537.

Lander E.S., Green P., Abrahamson J., Barlow A., Daly M.J., Lincoln S.E. and Newburg L. 1987. MAPMAKER APMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174–181.

Margiotti K., Kim E., Pearce C.L., Spera E., Novelli G. and Reichardt J.KV. 2002. Association of the G289S Single Nucleotide Polymorphism in the HSD17B3 Gene With Prostate Cancer in Italian Men. The Prostate 53: 65–68.

Oleykowski C.A., Mullins C.R.B., Godwin A.K., and Yeung A.T. 1998. Mutation detection using a novel plant endonuclease. Nucleic Acids Research 26: 4597–4602.

Precheur R.J. 2000. Ohio vegetable production guide. The Ohio State University Cooperative Extension, Bulletin 672.

Rozen S. and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Molecular Biology 132: 365–386.

Sacks E.J. and Francis D.M. 2001. Genetic and environmental variation for tomato flesh color in a population of modern breeding lines. J. Am. Soc. Hortic. Sci. 126: 221–226.

Smulders M.J.M, Bredemeijer G., Rus-Kortekaas W., Arens P. and Vosman B. 1997. Use of short microsatellites from database sequences to generate polymorphisms among Lycopersicon esculentum cultivars and accessions of other Lycopersicon species. Theor. Appl. Genet. 94: 264–272.

Sugimoto Y., Kuzushita N., Takehara T., Kanto T., Tatsumi T., Miyagi T., Jinushi M., Ohkawa K., Horimoto M., Kasahara A., Hori M., Sasaki Y. and Hayashi N. 2002. A single nucleotide polymorphism of the low molecular mass polypeptide 7 gene influences the interferon response in patients with chronic hepatitis C. Journal of Viral Hepatitis 9: 377–384.

Suliman-Pollatschek S., Kashkush K., Shats H., Hillel J. and Lavi U. 2002. Generation and mapping of AFLP, SSRs and SNPs in *Lycopersicon esculentum*. Cellular and Molecular Biology Letters 7: 583–597.

Tanksley S.D. and Jones R.A. 1981. Effects of O_2 stress on tomato alcohol dehydrogenase activity: description of a second ADH coding genes. Biochem. Genet. 19(3-4): 397–409.

Van der Hoeven R., Ronning C., Giovannoni J., Martin G. and Tanksley S.D. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14: 1441–1456.

Van der Knaap E. and Tanksley S.D. 2001. Identification and characterization of a novel locus controlling early fruit development in tomato. Theor. Appl. Genet. 103: 353–358.

Van Tuinen A., Cordonnier-Pratt M.M., Pratt L.H., Verkerk R., Zabel P. and Koornneef M. 1997. The mapping of phytochrome genes and photomorphogenic mutants of tomato. Theor. Appl. Genet. 94: 115–122.

Verhage B.A.J., van Houwelingen K., Ruijter T.E.G., Kiemeney L.A. and Schalken J.A. 2002. Single-nucleotidepolymorphisminth*ee-cadherin* gene promoter modifies the risk of prostate cancer. Int. J. Cancer 100: 683–685.

Yang B., Wen X., Kodali N.S., Oleykowdki C.A., Miller C.G., Kulinski J., Besack D., Yeung J.A., Kowalski D. and Yeung A.T. 2000. Purification, cloning, and characterization of the CEL I nuclease. Biochemistry 39: 3533–3541.

Yu J., Hu S.N., Wang J., et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science 296: 79–92.

Zhu Y.L., Hyatt S., Quigley C., Song Q.J., Grimm D., Young N. and Cregan P. 2001. Single nucleotide polymorphisms (snps) in soybean genes, cdnas, and random genomic sequence, *in* Plant & Animal Genome IX Conference, January 13-17, 2001. San Diego, California, USA.