cite this poster as:

Baldo, A.M., L.D. Robertson, and J.A. Labate. 2005. Highly polymorphic genes in cultivated tomato. HortScience 40:999.

Cultivated tomato varieties are genetically extremely similar. We identified 764 Unigenes with potential single nucleotide polymorphisms (SNPs) among more than 15 cultivars from public expressed tomato data. By sequencing regions from 53 of these Unigenes in two to three cultivars, we discovered an unexpected wealth of nucleotide polymorphism (62 SNPs and 12 indels in 21 Unigenes). This included a high proportion of predicted nonsynonymous nucleotide (17 of 33 SNPs in exons) and nonconservative amino acid (6 of 16 nonsynonymous SNPs) changes. We hypothesize that five of these regions are associated with introgressions from wild relatives. Identifying polymorphic, expressed genes in the tomato genome will be useful for both tomato improvement and germplasm conservation.

Poster Board # 6
Poster Session – 4
Genetics and Germplasm 1

Highly polymorphic genes in cultivated tomato*

*Manuscript accepted in Molecular Breeding Aug 5, 2005

Angela M. Baldo, Larry D. Robertson, and Joanne A. Labate USDA - ARS Plant Genetic Resources Unit, Geneva, NY 14456 http://www.ars-grin.gov/gen

Abstract

Cultivated tomato varieties are genetically extremely similar. We identified 764 Unigenes with potential single nucleotide polymorphisms (SNPs) among more than 15 cultivars from public expressed tomato data. By sequencing regions from 53 of these Unigenes in two to three cultivars, we discovered an unexpected wealth of nucleotide polymorphism (62 SNPs and 12 indels in 21 Unigenes). We hypothesize that five of these regions are associated with introgressions from wild relatives. Identifying polymorphic, expressed genes in the tomato genome will be useful for both tomato improvement and germplasm conservation.

Sixteen Conserved Ortholog Set (COS II) markers designed to span introns were also resequenced for SNPs in TA496, RioGrande, and Moneymaker.

Figure 1. Distribution of public tomato EST sequences

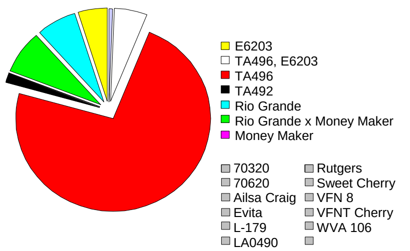
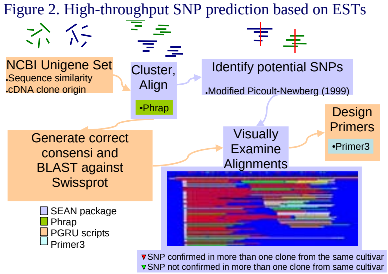


Figure 2. High-throughput SNP prediction based on ESTs



SNP Prediction

We have developed a data mining pipeline in PERL that screens an entire National Center for Biotechnology Information (NCBI) Unigene set (Wheeler et al. 2005) and provides an annotated list of predicted SNPs and PCR primers flanking them (Huntley et al. in prep). Our pipeline subclusters and aligns each Unigene using the SEAN SNP Prediction and Display Program (Huntley 2003). The consensus sequence of each subcluster was annotated using BLASTn against sequences of mapped markers in tomato from the Solanaceae Genomics Network (SGN) (Mueller et al 2004). SEAN, in turn, invokes Phrap (Green 2004). SEAN applies criteria designed to screen out potential sequencing errors (Picoult-Newberg et al. 1999). For a SNP to be called, there must be complete consensus among the alignment for seven nucleotides upstream of the SNP and seven downstream. Each SNP must be represented in at least two sequences. Using this method we identified 2,527 potential SNPs among 764 EST clusters from the NCBI tomato Unigene set.

Results

PCR primers were designed to amplify regions of predicted SNPs within 85 EST clusters. Eighty four primer pairs amplified fragments from genomic DNA that were resequenced in two or three cultivars predicted to contain SNPs. Roughly one-third of the regions amplified appeared to contain introns (Figure 3).

Fifty three primer pairs gave unambiguous DNA sequence data indicating whether or not SNPs were detected. The 31 remaining pairs either gave poor quality sequence, more than one PCR product, or insufficient data (Table 1). A total of 62 SNPs and 12 insertion-deletion (indel) polymorphisms were verified by two-pass sequencing within 21 of the 53 EST clusters (Table 3).

Table 1. Results of marker development for 85 tomato unigenes predicted to contain one or more SNPs.

	Number	Proportion†	Notes
Primer pairs tested	85		
Primer pairs that amplified	84	0.99	Discontinue testing the primer pair that hasn't amplified
Amplicon apparently contain intron	36	0.43	Observed amplicon was at least 50 bp larger than expected
Primer pair amplified more than one bad	9	0.11	Gel purify PCR amplicons and sequence to identify
Amplicon too large to sequence predicted SNP site, or sequence quality too poor to interpret	7	0.08	Redesign PCR primers based on genomic DNA sequence data that were collected
Only one cultivar sequenced	4	0.05	Test additional cultivars
Amplicon contained one or more SNPs	21	0.25	Test additional cultivars
Amplicon did not contain SNP	32	0.38	Test additional cultivars
SNP site was apparently heterozygous in sequenced cultivars	11	0.13	Generally many sites appeared heterozygous within cultivars, may not be good loci for marker development

† Based on 85 primer pairs for "Primer pairs that amplified", or 84 primer pairs for all other table rows.

References

Huntley, D. 2005. SEAN SNP prediction and display programs. (http://zebrafish.doc.ic.ac.uk/SEAN).

Miller J.C., and Tanksley S.D. 1990. RFLP Analysis of phylogenetic relationships and genetic variation in the genus *Lycopersicon*. Theor Appl Genet 80:437-448.

Mueller L.A., et. al. 2005. Solanaceae Genomics Network [Online] http://www.sgn.cornell.edu.

Nesbitt, T.C., and S.D. Tanksley. 2002. Comparative sequencing in the genus *Lycopersicon*: Implications for the evolution of fruit size in the domestication of cultivated tomatoes. Genetics 162:365-379.

Rozas J., Sánchez-DelBarrio J.C., et. al. R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.

Stevens M.A., and Rick C.M. 1986. Chapter 2. Genetics and breeding, p. 35-109, *In* Atherton J. and Rudich J. (eds), The tomato crop. Chapman and Hall, NY, NY.

Van der Beek J.G., et. al. 1992. Mapping strategy for resistance genes in tomato based on RFLPs between cultivars: *Cf9* (resistance to *Cladosporium fulvum*) on chromosome 1. Theor Appl Genet 84:106-112.

Wheeler DL, et. al. 2005. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D39-45.

Wu F, et. al. (2005). PCR-based Orthologous Gene Markers for Comparative Genomics and Phylogenetics in Asterid Plant Species. (in preparation) http://www.sgn.cornell.edu/markers/cosii_markers.html

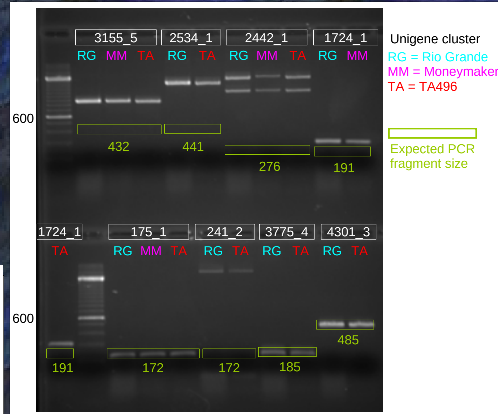


Figure 3.
Example of amplification results. Roughly one third of products were larger than expected.

Table 2. SNP frequencies and numbers of nucleotides sequenced for 53 primer pairs that gave unambiguous sequencing results†.

Class	No. of Primer Pairs	SNP Frequency [Nucleotides Sequenced] Exon + UTR	Intron	Total
One or more SNPs	21	0.0074 [2,572]	0.0050 [8,578]	0.0056 [11,150]
No apparent SNPs	32*	0.0000 [4,487]	0.0000 [5,192]	0.0000 [9,679]
Total	53	0.0027 [7,059]	0.0031 [5,192]	0.0030 [20,829]

† Two or three tomato cultivars were sequenced per primer pair.

* Resequencing 11 of these 32 in an expanded set of 30 landraces confirmed SNPs in 6 additional amplicons.

Table 3. Polymorphisms discovered among 21 tomato EST clusters by resequencing two or three cultivars

NCBI Unigene ID and subcluster	Chr[1], Marker	Swiss-Prot ID	E-value ‡	Description	dbSTS ID at GenBank[2]	Base pairs	Cultivar[3]	θ^4 all	θ^5 intron	Number of: SNPs Exon +UTR	Intron	Total	Indels Intron
2486_1					578995	240	T, E	0.0208	na	5		5	
2534_1R[6]	9, T0649	P23523	4 x 10-27	2-hydroxy-3-oxopropionate reductase (Tartronate semialdehyde reductase) (TASR)	578997	575	T, R	0.0157	0.0166		9	9	
437_2	1, 10, T0646				579009	662	T, R, M	0.0143	0.0169	3	6	9	9
220_1					578991	155	T, R	0.0129	na	2		2	
2325_3		P50160	5 x 10-49	Sex determination protein tasselseed 2	578992	413	T, R	0.0121	na	5		5	
3197_2					579004	147	T, R	0.0068	na	1		1	
3674_2					579008	148	T, R, M	0.0068	na	1		1	
1287_1					578988	151	T, R	0.0066	na	1		1	
175_1		O82528	4 x 10-80	60S ribosomal protein L15	578989	151	T, R, M	0.0066	na	1		1	
3081_1					579001	171	T, M	0.0059	na	1		1	
3332_3	7, T0643				579007	174	T, R	0.0057	na	1		1	
1909_2		P32495	1 x 10-35	High mobility group-like nuclear protein 2	578990	179	T, R	0.0056	na	1		1	
3284_1	1, T0214	P19446	1 x 10-142	Malate dehydrogenase, glyoxysomal precursor	579005	191	T, R	0.0053	na	1		1	
3155_3	5, cLET6I4	P35133	6 x 10-83	Ubiquitin-conjugating enzyme E2-17 kDa 10/12 (Ubiquitin-protein ligase 10/12) (Ubiquitin carrier protein 10/12)	579003	855	T, R, M	0.0047	0.0041	2	2	4	
2875_4					578998	1,366	T, R, M	0.0044	0.0047		6	6	1
3017_4		Q41951	3 x 10-66	Aquaporin TIP2.1 (Tonoplast intrinsic protein 2.1) (Delta-tonoplast intrinsic protein) (Delta-TIP)	579000	520	T, R	0.0039	0.0024	1	1	2	
2534_1F[6]	9, T0649				578996	601	T, R	0.0033	0.0000	2		2	2
241_2R[6]		P51298	9 x 10-27	Chloroplast 30S ribosomal protein S5	578994	649	T, R	0.0031	0.0031		2	2	
296_1					578999	979	T, R, M	0.0031	0.0012	2	1	3	
3132_3					579002	664	T, R	0.0030	0.0018	1	1	2	
1260_2	6, T0805	P31853	1 x 10-48	ATP synthase B' chain, chloroplast precursor (Subunit II)	578987	416	T, R, M	0.0024	na	1		1	
3300_2	4, CT188	P51074	3 x 10-87	Annexin-like protein RJ4	579006	550	T, M	0.0018	0.0000	1		1	
241_2F[6]					578993	659	T, R	0.0015	0.0017		1	1	
Total						10,616				33	29	62	12

[1] Loci with high confidence BLAST scores (E value = zero) to previously published, mapped DNA marker sequence.
[2] Primers, PCR protocols, thermoprofiles, and a representative amplicon sequence are available from dbSTS at NCBI
[3] Cultivars sequenced and seed sources were T = TA496 (Tanksley), E = E6203 (synonymous to TA209, Tanksley), R = Rio Grande (PI 303784) and M = Moneymaker (PI 286255)
[4, 5] Population diversity, $4N_e u$ = 4 x effective population size x mutation rate, based on the number of segregating sites between the two haplotypes for all sites (3) and intronic sites (4).
[6] Sequence of amplicon consists of two non-overlapping, forward (F) and reverse (R) segments.
‡ Only SwissProt matches with E-values < 10-20 are reported.

Discussion

Genetic bottlenecks, founder events, and selection have contributed to the uniformity of tomato (*Lycopersicon esculentum* cv. *esculentum*) (Miller and Tanksley, 1990). This lack of genetic diversity creates a challenge for characterizing crop germplasm collections and for continued improvement of cultivars. DNAsp (Rozas et al. 2003) was used to estimate population diversity, θ, for all sites and also separately for introns (Table 3).

Theta values for 20 of the sequences in Table 2 ranged from 1.8 to 6.8 SNPs per kb. These values were similar to random diversity found within *L. esculentum* var. *cerasiforme* (1.8 to 5.4 SNPs per kb, Nesbitt and Tanksley 2002 Table 4). Exons in Table 1 were generally more polymorphic than introns. This enriched polymorphism within exons reflects the fact that primers were designed to target an exonic SNP within a preferentially small (200 to 400 bp) amplicon.

These data lend preliminary support to the hypothesis that genetic variation in tomato cultivars is unevenly distributed, with rare islands of polymorphism that originated from introgression (van der Beek et al. 1992). In the early 1940s closely related wild species within the genus *Lycopersicon* were used as sources of disease resistance, and provided much of the breeding germplasm during subsequent decades (Stevens and Rick 1986).

There appear to be two classes of polymorphism values (0.0015 to 0.0068 versus 0.0121 to 0.0208), including results from introns (0.0017 to 0.0041 versus 0.0166 to 0.0169) in the data. The 0.0015 to 0.0068 range corresponds to θ estimates that have been observed within *L. esculentum* (0.0016 to 0.0054, Nesbitt and Tanksley 2002 Table 4). We hypothesize that the five most polymorphic regions in Table 3 (θ = 0.0121 to 0.0208) represent introgressions. Cultivars Rio Grande and Moneymaker are heirloom varieties collected in the early 1960s containing fewer introgressions than modern processing varieties TA496 and E6203. One of the hypothesized introgressed genes shares sequence similarity with maize Tassleseed-2, which has been implicated in its domestication from teosinte.

Conclusions

The SNP prediction rate over our 6.43 x 10^6 bp of computationally analyzed tomato consensus sequences was 3.93 x 10-4. Empirically we confirmed 28 of 103 predicted SNPs, yielding a transcriptome-wide estimate of 1.05 x 10-4 SNPs per nucleotide, i.e., 1 SNP per 9,542 nucleotides.

Our prediction method (62 SNP / 20829 bp sequenced, Table 2) yielded 27 times more confirmed polymorphism than the COS II markers (1 SNP / 9150 bp, data not shown) in the same three cultivars (TA496, RioGrande, Moneymaker). The COS II markers were designed to span introns in genes conserved between tomato and Arabidopsis. In such highly homogenous populations EST mining for polymorphism may be more fruitful than intron mining for TA496, RioGrande, and Moneymaker.

Acknowledgements

The authors would like to thank Derek Huntley for his continued collaboration with his SEAN SNP prediction package, Paul Kisly for his efforts in the greenhouse, Susan Sheffer and Katie Timmer for lab work, and also Katie for tables 1 & 2. We thank Feinan Wu and Steve Tanksley for the COS primer sequences. We would also like to thank the anonymous reviewers of this manuscript (in review) in the journal Molecular Breeding.