Cite this poster as:
Baldo, A.M., J. Labate, and L.D. Robertson. 2004. Prediction of single nucleotide polymorphisms in domestic tomato: How useful is EST sequence diversity? Intelligent Systems for Molecular Biology, Glasgow, UK.

Prediction of Single Nucleotide Polymorphisms in Domestic Tomato: How useful is EST sequence diversity?

Angela M. Baldo[1], Joanne Labate[2], Larry D. Robertson
[1]abaldo@pgru.ars.usda.gov, USDA-ARS Plant Genetic Resources Unit;
[2]jl265@cornell.edu, USDA-ARS Plant Genetic Resources Unit

Cultivated tomato is known to be relatively low in genetic diversity. This is a result of microevolutionary processes such as founder events, genetic bottlenecks, and intense selection. For these reasons, a computational approach to predicting single nucleotide polymorphisms (SNPs) is valuable to direct laboratory efforts toward regions more likely to yield results. We have developed a method to screen an entire NCBI Unigene set for potential SNPs using the SEAN SNP Prediction Program(Huntley, 2003). Predictions are based on established criteria: A window on either side of the predicted SNP must be identical for all sequences in the alignment, at least two sequences agreeing on each of a minimum of two polymorphisms, etc. Polymorphisms were further examined in the context of the cultivars and clones in which they were identified. Using this method, we discovered 2,527 potential SNPs among 764 clusters from the unigene set. We are in the process of verifying these polymorphisms in the laboratory, and comparing the results with sequence derived from randomly chosen introns, and diversity in the region of published, mapped markers.

Prediction of Single Nucleotide Polymorphisms in Domestic Tomato: How useful is EST sequence diversity?

Angela M. Baldo Joanne Labate and Larry D. Robertson

United States Department of Agriculture
Agricultural Research Service – Plant Genetic Resources Unit, Geneva, NY, USA
http://www.ars-grin.gov/gen

ABSTRACT

Cultivated tomato is known to be relatively low in genetic diversity. This is a result of microevolutionary processes such as founder events, genetic bottlenecks, and intense selection. For these reasons, a computational approach to predicting single nucleotide polymorphisms (SNPs) is valuable to direct laboratory efforts toward regions more likely to yield results. We have developed a method to screen an entire NCBI Unigene set for potential SNPs using the SEAN SNP Prediction Program (Huntley, 2003). Predictions are based on established criteria: A window on either side of the predicted SNP must be identical for all sequences in the alignment, at least two sequences agreeing on each of a minimum of two polymorphisms, etc. Polymorphisms were further examined in the context of the cultivars and clones in which they were identified. Using this method, we discovered 2,527 potential SNPs among 764 clusters from the unigene set. We are in the process of verifying these polymorphisms in the laboratory, and comparing the results with sequence derived from randomly chosen introns, and diversity in the region of published, mapped markers.

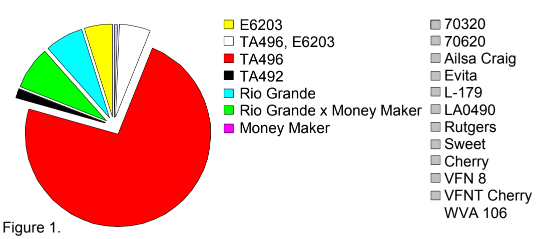
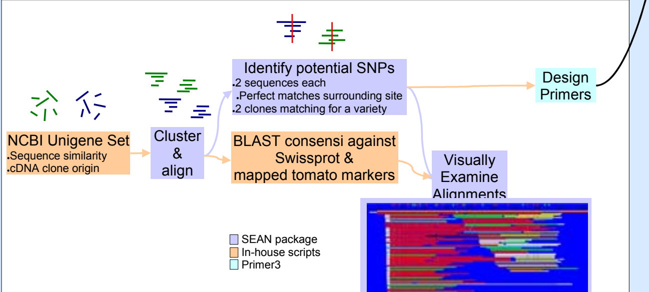


Figure 1.
Distribution of cultivars among publically available expressed tomato sequences

INTRODUCTION

The mission of PGRU is to characterize, distribute, and efficiently conserve large numbers of accessions for a variety of vegetable crop species and their wild relatives. DNA-based markers have been viewed as valuable tools to assist in genetic characterization of accessions. Vastly improved efficiency in DNA genotyping technology has made DNA markers an attractive and cost-effective method to gather precise genetic evidence in characterization of germplasm collections. The burgeoning amount of public sequence data in plants and emerging software tools greatly facilitates computational SNP prediction (Rafalski, 2002). In order to develop SNP markers in tomato we have leveraged the over 150,000 expressed tomato sequences in the public domain (specifically GenBank). There are a variety of cultivars represented in the data (Figure 1). With an estimated polymorphism frequency of one site in every 7KB (Nesbitt 2002), it was necessary to use computational methods where possible.


NCBI Unigene Set
- Sequence similarity
- cDNA clone origin

Cluster & align

Identify potential SNPs
- 2 sequences each
- Perfect matches surrounding site
- 2 clones matching for a variety

BLAST consensi against Swissprot & mapped tomato markers

Design Primers

Visually Examine Alignments

SEAN package
In-house scripts
Primer3
▼SNP confirmed in more than one clone from the same cultivar
▼SNP not confirmed in more than one clone from same cultivar

Figure2.
SNP Prediction and Confirmation

METHODOLOGY

We have created a data mining pipeline that takes an NCBI Unigene set and provides an annotated list of predicted SNPs and primers flanking them. The Unigene set compiled by NCBI is attractive because it includes data from all tomato ESTs deposited in the public domain, as well as the full-length cDNAs. Clusters are created using a variety of information (Pontius 2003). Our pipeline makes use of the SEAN SNP Prediction and Display Programs developed by Derek Huntley (2003) to sub-cluster and align each Unigene set by invoking Phrap (Green 1994). The SEAN package uses a variety of criteria to distinguish likely SNPs from sequence error, based on Picoult-Newberg (1999).

As a by-product of alignment a consensus sequence is produced for each contig. Our pipeline re-annotates the clusters, probing the consensus with BLASTx (Gish 1993) against the SwissProt database (Boeckmann 2003). The consensus is usually longer than the longest member of a UniGene set, and therefore often provides a stronger basis for similarity searching.

We visually examined the 764 clusters containing putative SNPs and selected 73 for testing. Primers were designed using Primer3 software (Rozen 2000) and used to amplify genomic DNA. Amplicons were directly sequenced for two or three of the cultivars represented in the cluster and the presence of a cultivar-specific SNP was determined.

RESULTS

Of the 73 attempted amplifications, 63 yielded a product. 23 of these reactions yielded a product larger than predicted (Figure 3). A few yielded two products, which were gel purified and sequenced separately. The magnitude of differences between expected and observed amplicon sizes ranged from 50 (the lowest difference detectable on our gels) to over 1300 nucleotides (Figure 4). These size differences appear to be due to the presence of introns. In no case was there a different size fragment between cultivars.

Of the 63 successful amplifications, 15 were either too large to sequence, or the sequence quality was too low to score. In 35 of the fragments the SNP was not confirmed, either because it was not found, or the position looked variable within the cultivar. As cultivated tomatoes are expected to be essentially homozygous, apparent within-individual variability may be due to multiple copy loci with low variability. A total of 12 cultivar-specific polymorphisms were confirmed. 11 of these matched the prediction, while one yielded an unexpected cultivar-specific difference.

Amplify & Sequence

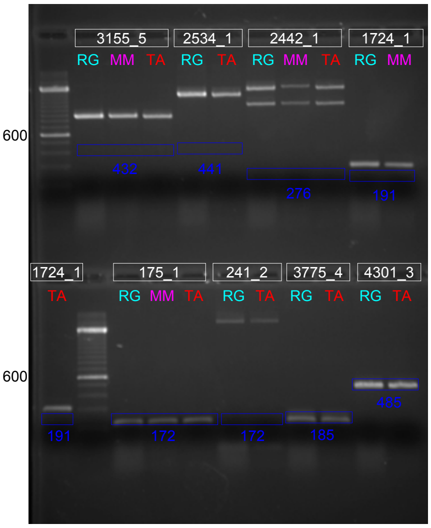


Figure 3.
Example of amplification results. Roughly one third of products were larger than expected. The difference in length ranged from 50 to over 1300 nucleotides (Figure 4).

73 Predictions Tested:

10 No Amplification
40 Observed = Expected Size
23 Observed > Expected Size

25 Prediction was wrong
10 Looks heterozygous
12 Poor Quality Sequence
1 Doesn't match consensus
3 Too Large to sequence
11 Confirmed
1 Unexpected SNP

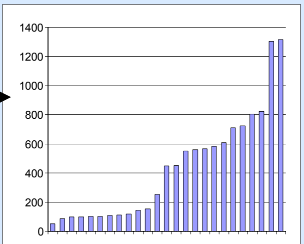


Figure 4.
Amplicons longer than expected

Unigene_seg	Expected Amplicon	Observed â€ Expected size	Best Swissprot Hit	E-Value	
1260_2	416	0	ATP synthase B' chain, chloroplast precursor (S	1.00E-048	
1287_1	195	0	Integrin beta-5 precursor	0	
1909_2	200	0	High mobility group-like nuclear protein 2	1.00E-035	
2325_3	459	0	SEX DETERMINATION PROTEIN TASSELSE	4.00E-049	
2486_1	253	0	Ancient ubiquitous protein 1 precursor	0.23	
3081_1	198	0	Acetylornithine aminotransferase (ACOAT)	2.2	
3284_1	186	0	Malate dehydrogenase, glyoxysomal precursor	1.00E-142	
3302_2	260	0	26S protease regulatory subunit 8 (Proteasome	1.00E-173	
3300_2	417	154	Annexin-like protein RJ4	3.00E-087	
3132_3	151	549	Ribulose bisphosphate carboxylase/oxygenase	0.05	
296_1	176	824	Chloride intracellular channel 6	4.00E-016	
2875_4	197	1303	60S ribosomal protein L18a	3.00E-092	(unexpectected SNP)

DISCUSSION

Our method yielded 12 cultivar-specific SNPs, which were confirmed by sequencing 12.782 KB. This is equivalent to a rate of one SNP per 1065 BP, roughly six times more frequent than might be expected from raw estimates of polymorphism in tomatoes (Nesbitt 2002). Only one of the 12 polymorphisms occurs in an intron. An additional 16 of our predicted SNPs (not shown here) agree with a set of 100 reported by Yang & Francis (in press) for the purposes of mapping. One of those 16 appears to be polymorphic within TA496.

ACKNOWLEDGEMENTS

We are very grateful to the developers of the software packages used, particularly Derek Huntley, for our frequent discussions and his enthusiasm for incorporating our suggestions. The authors also thank Katie Timmer and Susan Sheffer for their heroic laboratory work and visual inspection of the SNP predictions, and Mike D'Amico for the primer predictions. This project was funded by the USDA-ARS, CRIS project #1907-21000-016-00.

REFERENCES

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in Nucleic Acids Res. 31:365-370.
Gish, W. & States, D.J. 1993. "Identification of protein coding regions by database similarity search." Nature Genet. 3:266-272.
Green, P. 1994. phrap (http://www.phrap.org)
Huntley, D. 2003. SEAN SNP prediction and display programs. (http://zebrafish.doc.ic.ac.uk/SEAN).
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.
Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M 1999. Mining SNPs from EST databases. Genome Res 9: 167174
Pontius JU, Wagner L, Schuler GD. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information; 2003.
Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Current Opinion in Plant Biology 5:94-100.
Rozen, S., Skaletsky, H. "Primer3 on the WWW for general users and for biologist programmers." In S. Krawetz and S. Misener, eds. Bioinformatics Methods and Protocols in the series Methods in Molecular Biology. Humana Press, Totowa, NJ, 2000, pages 365-386.
Yang, W, X. Bai, E. Kabelka, C. Eaton, S. Kamoun, E. van der Knaap, and D. Francis. (in press) Discovery of single nucleotide polymorphisms in Lycopersicon esculentum by computer aided analysis of expressed sequence tags. Molecular Breeding.