Cite this poster as:

**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 donors 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 SNPs and primers flanking them. The Unigene set compiled by NCBI is attractive because it includes randomly chosen introns, and diversity in the region of published, mapped markers.

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 processing 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, 2003). In order to develop SNP markers in tomato we have leveraged the over 150,000 expressed sequence tags 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 (Nessbit 2002), it was necessary to use computational methods where possible.

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.

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 difference 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.

The consensus is usually longer than the longest member of a UniGene set, and therefore often provides a stronger basis for similarity searching.

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 Shaffer 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


Role, S. and Moussian, B. "Phytochae, the "path the UNIPRO for general uses and the Goltic program". http://www.upse.upse.35-35/35/moussian/bio/Phytochae.html (in press).

Role, S. and Moussian, B. "Phytochae, the "path the UNIPRO for general uses and the Goltic program". http://www.upse.upse.35-35/35/moussian/bio/Phytochae.html (in press).
