



## Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP

Gary Barker<sup>1</sup>, Jacqueline Batley<sup>2</sup>, Helen O' Sullivan<sup>1</sup>,  
Keith J. Edwards<sup>3</sup> and David Edwards<sup>2,\*</sup>

<sup>1</sup>Institute of Arable Crop Research, Long Ashton, Bristol, BS41 9AF, UK, <sup>2</sup>Agriculture Victoria Plant Biotechnology Centre, La Trobe University, Bundoora, Victoria 3086, Australia and <sup>3</sup>School of Biological Sciences, University of Bristol BS8 1UG, UK

Received on June 28, 2002; revised on August 23, 2002; accepted on August 28, 2002

### ABSTRACT

**Summary:** AutoSNP is a program to detect single nucleotide polymorphisms (SNPs) and insertion/deletion polymorphisms (indels) in expressed sequence tag (EST) data. The program uses d2cluster and cap3 to cluster and align EST sequences, and uses redundancy to differentiate between candidate SNPs and sequence errors. Candidate polymorphisms are identified as occurring in multiple reads within an alignment. For each candidate SNP, two measures of confidence are calculated, the redundancy of the polymorphism at a SNP locus and the co segregation of the candidate SNP with other SNPs in the alignment.

**Availability:** The program was written in PERL and is freely available to non-commercial users by request from the authors.

**Contact:** dave.edwards@nre.vic.gov.au

Single nucleotide polymorphisms (SNPs) are increasingly becoming the marker of choice in genetic analysis. They are used routinely in agriculture as markers in breeding programs and have many uses in human genetics, such as the detection of alleles associated with genetic diseases and the identification of individuals. SNPs are invaluable for genome mapping, offering the potential for generating very high density genetic maps (Rafalski, 2002). The low mutation rate of SNPs also makes them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (Syvanen, 2001).

As with the majority of molecular markers, one of the limitations of SNPs is the initial cost associated with their development. However, with the development of high throughput sequencing technology, large amounts of data have been submitted to the various DNA databases that may be suitable for data mining and SNP discovery

(Taillon-Miller *et al.*, 1998). In particular, EST sequencing programs have provided a wealth of information, identifying novel genes from a broad range of organisms. EST sequence data may provide the richest source of biologically useful SNPs due to the relatively high redundancy of gene sequence, the diversity of genotypes represented within databases and the fact that each SNP would be associated with an expressed gene. Methods used to identify SNPs in aligned sequence data has previously relied on sequence trace file analysis to filter out sequence errors by their dubious trace quality (Kwok *et al.*, 1994; Marth *et al.*, 1999; Garg *et al.*, 1999). The major drawbacks to this approach are the requirement for sequence trace files, which are rarely complete for large sequence datasets collated from a variety sources, and the high level of sequence error associated with the reverse transcription process.

We have attempted to overcome this difficulty by developing software for the automated detection of SNPs within EST data with associated measurements of confidence in the validity of candidate SNPs. A conservative approach was followed to limit the error associated with cloning and sequencing, so that only polymorphisms represented by two or more sequences were considered. While this discards a significant amount of variation in the EST data, it permits the ready identification of large numbers of candidate SNPs with a high level of confidence in their validity.

### PROGRAM OPTIONS

The AutoSNP script is run from the command line. On start-up, the user is asked to supply FASTA format input file name together with a similarity cut-off for d2cluster and cap3. Default values are 80% similarity for d2cluster and 95% for cap3.

### PROGRAM FLOW AND DEPENDENCIES

Initial clustering is carried out by d2cluster (Burke *et al.*, 1999). AutoSNP reads the output table created by

\*To whom correspondence should be addressed.

Key to sequences:

A B73 A1855425  
 B B73 A1964458  
 C B73 A1964454  
 D B73 A1964498  
 E W23 AW288875  
 F W23 AW288876  
 G W23 AW289056  
 H W23 AW331212  
 I W23 AW787314  
 J W23 AW787315  
 K W23 AW787332  
 L W23 BB025302  
 M W23 BB025303  
 N W23 BE129644  
 O W23 BE129897  
 P W23 BE225008  
 Q W23 BE519299  
 R BPT29420  
 S BG840383  
 T AA979839  
 U A1372183

base	ABCDEFGHIJKLMNQRSTU	Min. informative	Cosegregation	Weighted cosegregation
34	G...AA.....A.....GGA	3	9/11	27
48	AAA.GG.GG.GG...AAG	5	9/11	55
154	AAA.CCCCCCCCCC.AAC	5	9/11	74
230	CCC.CCCCTCCCCCCCCC	2	1/11	8
242	CCC.....CC	5	9/11	74
243	CCC.....CC	5	9/11	74
244	AAA.....AA	5	9/11	74
295	GGG.AAAAAAAAAAAGG	5	9/11	70
323	GGG.TTTTTTTTTTTGG	5	9/11	70
400	CCCC.TTTTTTTTTTCC	5	9/11	74
440	TTTT.TTTTTT.TTCC.TT	2	1/11	7

**Fig. 1.** An AutoSNP report summary. This report depicts 11 candidate SNPs, identifying their base position in the sequence alignment along with two measures of confidence in SNP validity. The Min. informative score measures the minimum number of sequences that represent a polymorphism. The cosegregation score is a measure of the number of SNPs in the alignment which share the same pattern of polymorphism between aligned sequences. The weighted cosegregation score takes account of missing data in the alignment of ESTs that may otherwise bias the cosegregation score. The key relates the aligned sequences to original GenBank sequence and also identifies the maize line (where available) derived from the GenBank annotation. The full SNP report includes the complete sequence alignment along with the above SNP summary.

d2cluster, and uses the information to build sequence cluster files in FASTA format. These clusters are then passed to the sequence assembly program cap3 (Huang and Madan, 1999). AutoSNP reads the ACE format output file from each cap3 run, and generates gapped FASTA format alignment files which are finally passed to the SNP detection and co-segregation subroutines.

## PROGRAM OUTPUT

The primary output of AutoSNP is a set of linked HTML format SNP reports, prefaced by an index page containing statistical information relating to the sequence contig assembly and candidate SNP/indel identification. The SNP report pages have three components: (i) A key to the sequences in the alignment, (ii) A summary table showing the candidate SNPs/indels, together with confidence scores, and (iii) A full vertical alignment of the sequences, with the SNPs highlighted (Figure 1). Each SNP report also has a hyperlink to the underlying sequence alignment in FASTA format. In addition to the main report, several supporting files are produced which hold information such as the frequency distribution of cap3 sequence contig sizes, and the number of SNPs associated with each size of sequence contig, nucleotide substitution ratios and tables of indel sequence and size frequency.

## PERFORMANCE WITH THE MAIZE TEST DATA

An input file containing 102 551 maize ESTs was downloaded from ZmDB (<http://www.zmdb.iastate.edu/>), and the AutoSNP program executed on a 1 GHz Intel Pentium III PC with 520 MB RAM running RedHat Linux 7.0. The d2cluster program took 6 days to organize the sequences into primary clusters. The cap3 assembly and SNP detection took a further 22 h to complete analysis. Of the 13 247 clusters produced by cap3, 3479 were found to contain one or more candidate SNP. A total of 14832 candidate polymorphisms were identified (<http://www.cerealsdb.uk.net/discover.htm>). Indel size frequencies, nucleotide substitution ratios and segregation of candidate polymorphisms with haplotypes indicate that the majority of SNPs and indels identified using this approach represent true genetic variation in maize.

## ACKNOWLEDGEMENTS

IACR-Long Ashton receives grant aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. David Edwards and Gary Barker were funded under the BBSRC IGF initiative (IGF12403), Helen O'Sullivan is funded by BBSRC grant D14009.

## REFERENCES

- Burke,J., Davison,D. and Hide,W. (1999) d2\_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.*, **9**, 1135–1142.
- Garg,K., Green,P. and Nickerson,D.A. (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res.*, **9**, 1087–1092.
- Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
- Kwok,P.Y., Carlson,C., Yager,T.D., Ankeney,W. and Nickerson,D.A. (1994) Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics*, **23**, 138–144.
- Marth,G.T., Korf,I., Yandell,M.D., Yeh,R.T., Gu,Z.J., Zakeri,H., Stitzel,N.O., Hillier,L., Kwok,P.Y. and Gish,W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genet.*, **23**, 452–456.
- Pearson,W.R. (1990) Rapid and Sensitive Sequence Comparison with FASTP and FASTA. *Method Enzymol.*, **183**, 63–98.
- Rafalski,A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.*, **5**, 94–100.
- Syvanen,A.C. (2001) Accessing genetic variation Genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.*, **2**, 930–942.
- Taillon-Miller,P., Gu,Z.J., Li,Q., Hillier,L. and Kwok,P.Y. (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.*, **8**, 748–754.