Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W493–W495***
doi:10.1093/nar/gki462

SNPServer: a real-time SNP discovery tool

David Savage[1], Jacqueline Batley[1], Tim Erwin[1,2], Erica Logan[1,2], Christopher G. Love[1,2], Geraldine A. C. Lim[1,2], Emmanuel Mongin[1,2], Gary Barker[3], German C. Spangenberg[1,2] and David Edwards[1,2,*]

[1]Plant Biotechnology Centre and [2]Victorian Bioinformatics Consortium, Plant Biotechnology Centre, Primary Industries Research Victoria, La Trobe University, Bundoora 3086, Victoria, Australia and [3]School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK

Received February 11, 2005; Revised March 23, 2005; Accepted April 13, 2005

ABSTRACT

SNPServer is a real-time flexible tool for the discovery of SNPs (single nucleotide polymorphisms) within DNA sequence data. The program uses BLAST, to identify related sequences, and CAP3, to cluster and align these sequences. The alignments are parsed to the SNP discovery software autoSNP, a program that detects SNPs and insertion/deletion polymorphisms (indels). Alternatively, lists of related sequences or pre-assembled sequences may be entered for SNP discovery. SNPServer and autoSNP use redundancy to differentiate between candidate SNPs and sequence errors. 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. SNPServer is available at http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) and small insertions/deletions (indels) are the most frequently found DNA sequence variations (1). The development of high-throughput methods for the detection of SNPs has led to a revolution in their use as molecular markers (2). As such, they represent one of the most powerful tools for the analysis of genomes and are increasingly becoming the marker of choice in genetic analysis. SNPs are used routinely in agriculture as markers in crop-breeding programmes (3). They also have many uses in human genetics, such as the detection of alleles associated with genetic diseases and inferences of population history (4,5). Furthermore, SNPs are invaluable as a tool for genome mapping, offering the potential for generating very high-density genetic maps, that can be used to develop haplotyping systems for genes or regions of interest (6). The simplicity and the low mutation rate of SNPs also make them excellent markers for studying complex genetic traits and as a tool for the understanding of genome evolution (7).

As with the majority of molecular markers, one of the limitations of SNPs is the initial cost associated with their development. However, with the growth 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 (8). Methods used to identify SNPs in aligned sequence data have traditionally relied on sequence trace file analysis to filter out sequence errors by their dubious trace quality (9–11). 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 of sources, and the high level of sequence error associated with the reverse transcription process. These problems are overcome by the use of autoSNP software for the detection of SNPs within sequence data with associated measurements of confidence (12).

AutoSNP calculates two associated measurements of confidence in the validity of SNPs for each polymorphism. The frequency of occurrence of a polymorphism at a particular locus provides a primary measure of confidence in the SNP representing a true polymorphism and is referred to as the SNP redundancy score. The co-segregation of multiple SNPs within an alignment to define a haplotype provides a second measure of confidence in SNP validity and is referred to as the co-segregation score. Here we introduce the real-time autoSNP web server, the SNPServer. This builds on the use of autoSNP software by providing a web interface for sequence input, comparison and assembly, and permits the rapid discovery of SNPs related to any specified sequence of interest.

*To whom correspondence should be addressed. Tel: +61 3 9479 5633; Fax: +61 3 9479 3618; Email: Dave.Edwards@dpi.vic.gov.au

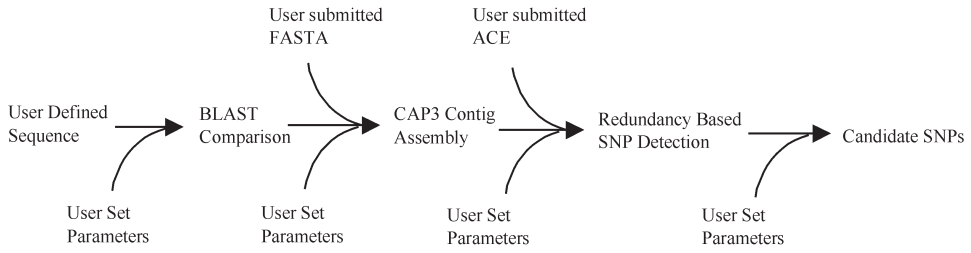

Figure 1. An overview of components of the real-time autoSNP web server, the SNPServer.

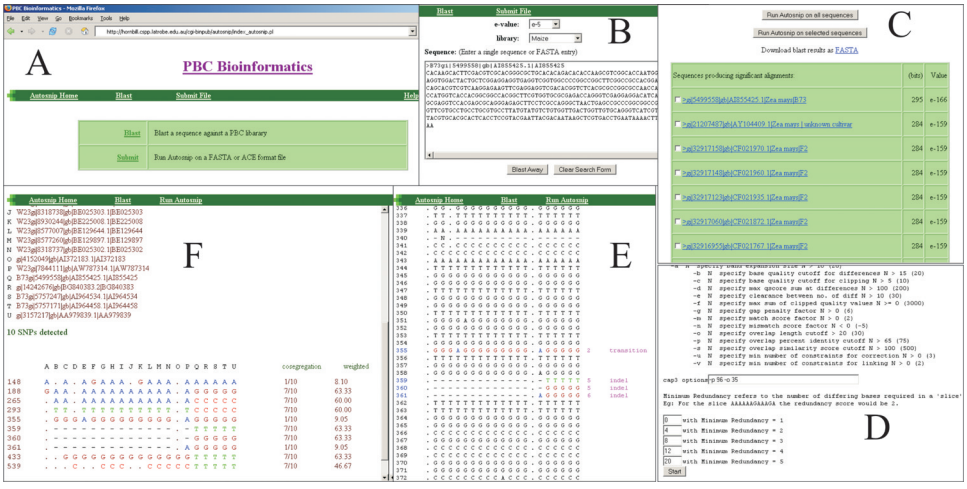

Figure 2. On entry of a sequence and specification of a sequence database for comparison (**A** and **B**), a BLAST job is initiated. On completion, a summary of matching sequences are displayed permitting the selection of sequences for assembly (**C**). The assembly page permits users to specify CAP3 and autoSNP parameters (**D**). A summary page provides information on the assembly and the SNP discovery, and permits users to return to the assembly page to modify parameters. The results page consists of two windows, the first provides a complete vertical alignment, highlighting SNPs (**E**) while the second lists the assembly member sequences and provides a SNP summary (**F**).

METHODS

Sequence input, assembly and clustering

The real-time autoSNP web server, SNPServer, acts as a web interface and wrapper for the three programs, BLAST, CAP3 and autoSNP, that make up the SNP discovery pipeline (Figure 1). The complete pipeline accepts a single sequence as an input. This entry sequence is compared with a specified nucleotide sequence database using BLAST (13) to identify related sequences. The resulting sequences may then be selected for assembly with CAP3 (14) and subsequent SNP discovery using autoSNP (12). Alternatively, users may enter a list of sequences in FASTA format for assembly, or a pre-calculated sequence assembly in ACE format. Complete options for BLAST sequence comparisons, CAP3 assembly and SNP discovery may be specified at the user interface.

SNP discovery

SNP discovery is performed using a redundancy-based approach with a modified version of the autoSNP PERL script (12,15). Alignment data generated by CAP3 (or from a user submitted ACE file) are used to load the sequences in each assembly into a 2D array. Spacing characters (-) added during sequence alignment are considered as a fifth element in addition to the four nucleotides A, C, G and T. This permits the identification of insertion/deletion polymorphisms between sequences. Each row (representing a single base locus in the assembly) is assessed for differing nucleotides. Minimum redundancy scores specified by the user and associated with alignment width (the number of sequences included in the contig) determine the number of different nucleotides at a base position required for classification as a SNP. Where a SNP is recorded, an SNP score is allocated equal to the minimum number of reads that share a common polymorphism. Where several SNPs are present in an alignment, a co-segregation score is calculated for each SNP. This is measured as the frequency of haplotype specifying SNP patterns occurring in the alignment. This figure is then normalized to the number of sequences in the alignment to produce a weighted co-segregation score. HTML format files are generated to allow the user to input data, select comparison, assembly and SNP discovery parameters, and browse the SNP results (Figure 2).

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the Victorian Department of Primary Industries.

Conflict of interest statement. None declared.

REFERENCES

1. Kwok,P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Ann. Rev. Genom. Hum. Genet*., **2**, 235–258.
2. Twyman,R.M. and Primrose,S.B. (2003) Techniques patents for SNP genotyping. *Pharmacogenomics*, **4**, 67–79.
3. Gupta,P.K., Roy,J.K. and Prasad,M. (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci*., **80**, 524–535.
4. Brumfield,R.T., Beerli,P., Nickerson,D.A. and Edwards,S.V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol*., **18**, 249–256.
5. Collins,A., Lau,W. and De la Vega,F.M. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum. Hered*., **58**, 2–9.
6. Rafalski,A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol*., **5**, 94–100.
7. Syvanen,A.C. (2001) Accessing genetic variation. Genotyping single nucleotide polymorphisms. *Nature Rev. Genet*., **2**, 930–942.
8. 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.
9. 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.
10. Kwok,P.Y., Carlson,C., Yager,T.D., Ankener,W. and Nickerson,D.A. (1994) Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics*, **23**, 138–144.
11. Marth,G.T., Korf,I., Yandell,M.D., Yeh,R.T., Gu,Z.J., Zakeri,H., Stitziel,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.
12. Barker,G., Batley,J., O'Sullivan,H., Edwards,K.J. and Edwards,D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics*, **19**, 421–422.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
14. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res*., **9**, 868–877.
15. Batley,J., Barker,G., O'Sullivan,H., Edwards,K.J. and Edwards,D. (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol*., **132**, 84–91.