Sequence analysis

SEAN: SNP prediction and display program utilizing EST sequence clusters

Derek Huntley1,*, Angela Baldo4, Saurabh Johri2 and Marek Sergot3

1Centre for Bioinformatics, Division of Molecular Biosciences, 2Centre for Molecular Microbiology and Infection, Division of Investigative Sciences and 3Department of Computing, Imperial College, London SW7 2AZ, UK and 4USDA-ARS Plant Genetic Resources Unit, New York State Agricultural Experiment Station, Geneva, NY 14456, USA

Received on June 29, 2005; revised on November 23, 2005; accepted on December 11, 2005

Advance Access publication December 15, 2005

Associate Editor: Chris Stoeckert

ABSTRACT

Summary: SEAN is an application that predicts single nucleotide polymorphisms (SNPs) using multiple sequence alignments produced from expressed sequence tag (EST) clusters. The algorithm uses rules of sequence identity and SNP abundance to determine the quality of the prediction. A Java viewer is provided to display the EST alignments and predicted SNPs.

Availability: SEAN is freely available from http://zebrafish.doc.ic.ac.uk/Sean

Contact: d.huntley@imperial.ac.uk

INTRODUCTION

Expressed sequence tags (ESTs) are an important resource for identifying polymorphisms in transcribed regions. In humans, for example, estimates of polymorphism are in the range of 1 every 1.3 kb (Sachidanandam et al., 2001) and in cultivated tomatoes 1 every 7 kb (Nesbitt and Tanksley, 2002). SEAN provides a method to predict and visualize the presence of single nucleotide polymorphisms (SNPs) using EST sequence clusters. EST data have previously been used for SNP prediction by programs such as AutoSNP (Barker et al., 2003), PolyPhred (Nickerson et al., 1997), PolyBayes (Marth et al., 1999), TRACE_DIFF (Bonfield et al., 1998) and HarvEST (HarvEST Home Page available at http://harvest.ucr.edu). Whereas HarvEST provides pre-built SNP prediction libraries, AutoSNP, PolyPhred and PolyBayes, like SEAN, enable the prediction of SNPs from a users own EST dataset. SEAN, as with AutoSNP, uses the redundancy of the SNP in an alignment as a measure of confidence but reinforces this with a measure of sequence identity in the surrounding aligned sequences. Unlike the other tools listed, SEAN also allows for the inclusion of library data to further support SNP predictions. A Java viewer is included that enables the visualization of the alignments and SNP predictions for user inspection.

The search strategy for SEAN is based on the work of Picoult-Newberg et al. (1999) The sequence assembly program Phrap (Phrap available at http://www.phrap.org) is used to build a consensus from the clustered sequences and using the output file produced by the Phrap ‘ace’ flag the sequence alignment, including consensus, is built and the alignment parsed to find potential SNPs.

Five output files are produced by SEAN: three reference files and two Java configuration files. The first two reference files contain the sequence alignments (only those regions that align with the consensus are in the first file, the full alignments are in the second) together with a list of the potential SNPs and their locations and the consensus sequence in FASTA format. The third reference file contains a listing of the contigs produced by Phrap and their details—sequences, average sequence length and number of predicted SNPs. There is an option to include cultivar and library information for an improved SNP prediction. If this is used an additional output file details the predicted SNP position within the consensus and the number of occurrences of each base within each library at that position. This is provided to give additional evidence of the quality of the predicted SNP.

There are also two Java configuration files produced, one for the alignments only and one for the complete sequences. These are for a Java viewer that has been developed to enable visual inspection of the alignments and predicted SNPs. The viewer has been developed using the Neomorphic Genomic Software Development Kit (NGSDK) (available at http://www.affymetrix.com). The viewer displays the sequences as solid bars with the position of any potential SNPs shown by red points at the top of the display and the positions in the relevant sequences highlighted in red. If the SNP predictions have been generated using the library and cultivar data then the SNPs predicted with the lower confidence are coloured green to distinguish them. The display has zooming functionality and fully horizontally zooming overlays the bars with their nucleotide sequence.

SEAN requires Perl and Phrap for the analysis component and Java (1.3+) for the viewer.

IMPLEMENTATION

The SEAN generated sequence alignment is parsed a base position at a time to find potential SNPs by comparing the base at each position with the corresponding consensus base. To eliminate poor quality sequence when a base difference is found, the surrounding sequence is compared with the consensus over a defined window, by default 15 bp either side of the base but configurable when running SEAN. If the sequences in the windows are identical the base and its position are flagged, and stored as a predicted SNP.
only if another identical base change is found at the same position in at least one other sequence. A further check is also made that the consensus base is also present in at least two sequences, as the consensus produced by Phrap does not always contain the dominant base at a particular position.

The prediction requirements mean that for a potential SNP to be found at least four overlapping sequences are required. Clusters containing large numbers of sequences are also unusable owing to the memory requirements of Phrap. The resources of the host computer determine the limit; on a standard PC with 1 Gb RAM up to 500 sequences can be handled satisfactorily, depending upon their compositional similarity. Pre-clustering sequences using Cap3 (Huang and Madan, 1999) or Gap (Bonfield et al., 1995) can reduce the number of sequences handled simultaneously.

If the sequences within the window either side of the potential SNP contain gaps in order to facilitate alignment, the window size is increased accordingly so that the actual required number of nucleotides are checked. Gaps are also sometimes included in the consensus produced by Phrap when they are not in the majority of the aligned sequences, occasionally when present in only one aligned sequence. If such gaps are present within the window region it could prejudice against the selection of a potential SNP so to compensate for this the window sequences are screened to ensure they are identical at the associated position.

The quality of the SNP predictions can be strengthened by the inclusion of cultivar and library data. SEAN reads an optional file containing this data for each sequence and separately labels potential SNPs that are predicted where they are present in at least two libraries from the same cultivar. These SNPs are also coloured differently in the Java viewer so that they can be readily identified.

VALIDATION

In silico validation of SEAN has been carried out by searching mouse and human UniGene (Boguski and Schuler, 1995) clusters and confirming the predicted SNPs using the relevant dbSNP databases (Sherry et al., 2001). UniGene clusters were selected with the minimum number of four sequences required for SNP prediction and a maximum number of 500. This provided 27 169 human and 29 360 mouse clusters from which 128 408 human and 328 714 mouse SNPs were predicted. dbSNP contained 9 123 517 human and 29 360 mouse clusters from which 128 408 human and 328 714 mouse SNPs were predicted. dbSNP contained 9 123 517 human and 29 360 mouse clusters from which 128 408 human and 328 714 mouse SNPs were predicted. dbSNP contained 9 123 517 human and 29 360 mouse clusters from which 128 408 human and 328 714 mouse SNPs were predicted.

SEAN has been used to successfully identify SNPs among public ESTs from tomato cultivars. Among 53 re-sequenced contigs in two or three cultivars, 21 confirmed the SNPs predicted by SEAN (Labate and Baldo, 2005). Five additional SNPs were visible in the SEAN viewer but not predicted because they fell within 15 bp of each other. Overall efficiency of SNP discovery/confirmation was increased 10-fold using SEAN to target SNP-containing regions relative to sequencing arbitrary regions of the genome (Labate and Baldo, 2005). Further validation results are documented on the website (SEAN SNP prediction and display programs available at http://zebrafish.doc.ic.ac.uk/Sean/).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Joanne Labate for the confirmation data of SNPs in cultivated tomato and constructive suggestions for improvements in the SEAN prediction package and viewer. The authors also thank Elizabeth Fisher for the original idea for the development of SEAN.

Conflict of Interest: none declared.

REFERENCES