Published online 14 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D951–D953***
doi:10.1093/nar/gkn650

AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants

Chris Duran[1,2], Nikki Appleby[1,2], Terry Clark[1,2], David Wood[1,3], Michael Imelfort[1,2], Jacqueline Batley[1,4] and David Edwards[1,2,*]

[1]Australian Centre for Plant Functional Genomics, Brisbane, School of Land, Crop and Food Sciences, [2]The Institute for Molecular Bioscience, [3]Queensland Facility for Advanced Bioinformatics, The Institute for Molecular Bioscience, ARC Centre of Excellence in Bioinformatics and [4]ARC Centre of Excellence for Integrative Legume Research, University of Queensland, Brisbane, QLD 4072, Australia

Received August 12, 2008; Revised September 15, 2008; Accepted September 18, 2008

ABSTRACT

Single nucleotide polymorphisms (SNPs) may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence (a single nucleotide), are generally abundant in populations and have a low mutation rate. Analysis of assembled EST sequence data provides a cost-effective means to identify large numbers of SNPs associated with functional genes. We have developed an integrated SNP discovery pipeline, which identifies SNPs from assembled EST sequences. The results are maintained in a custom relational database along with EST source and annotation information. The current database hosts data for the important crops rice, barley and Brassica. Users may rapidly identify polymorphic sequences of interest through BLAST sequence comparison, keyword searches of annotations derived from UniRef90 and GenBank comparisons, GO annotations or in genes corresponding to syntenic regions of reference genomes. In addition, SNPs between specific varieties may be identified for targeted mapping and association studies. SNPs are viewed using a user-friendly graphical interface. The database is freely accessible at http://autosnpdb.qfab.org.au/.

INTRODUCTION

Molecular genetic markers describe genetic variations and provide a link between observed phenotypes and the underlying genotype. The development of high-throughput methods for the detection of single nucleotide polymorphisms (SNPs) and small insertion/deletions (indels) has led to a revolution in their use as molecular markers. SNPs may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence, are generally abundant in populations and have a low mutation rate (1). However, SNP markers can be costly to develop, especially where re-sequencing from multiple individuals is required. The mining of readily available sequence data significantly reduces the costs associated with SNP discovery (2). The principal challenge in SNP discovery remains the discrimination between true genetic polymorphisms and the often more abundant sequence errors. Where sequence trace files are available to filter polymorphisms in traces of dubious quality, these can be used to differentiate between true SNPs and sequence error (3). Where trace files are unavailable, the identification of sequence errors can be based on two further methods to determine SNP confidence: redundancy of the polymorphism in a sequence alignment, and co-segregation of putative SNPs with haplotype. The frequency of occurrence of a polymorphism at a particular locus provides a measure of confidence that the SNP represents a true polymorphism; this is referred to as the SNP redundancy score. In addition, true SNPs that represent divergence between homologous genes co-segregate to define a conserved haplotype. A co-segregation score based on whether a SNP position contributes to defining a haplotype provides a second independent measure of SNP confidence. The SNP redundancy score and co-segregation score together provide an effective means for estimating confidence in the validity of SNPs independently of sequence trace files (4–6).

We have combined SNP discovery software and sequence annotation within the relational database schema of autoSNPdb to enable the efficient identification of SNP and indel polymorphisms related to specific genes or traits. Here, we present the application of autoSNPdb to barley, rice and *Brassica* species. AutoSNPdb has a flexible interface facilitating a variety of queries. Users may search for SNPs within genes of predicted function, and through sequence identity with known genes. In addition, it is possible to add additional levels of annotation

*To whom correspondence should be addressed. Tel: +61 (0)7 3346 2615; Fax: +61 (0)7 3346 2101; Email: Dave.Edwards@uq.edu.au

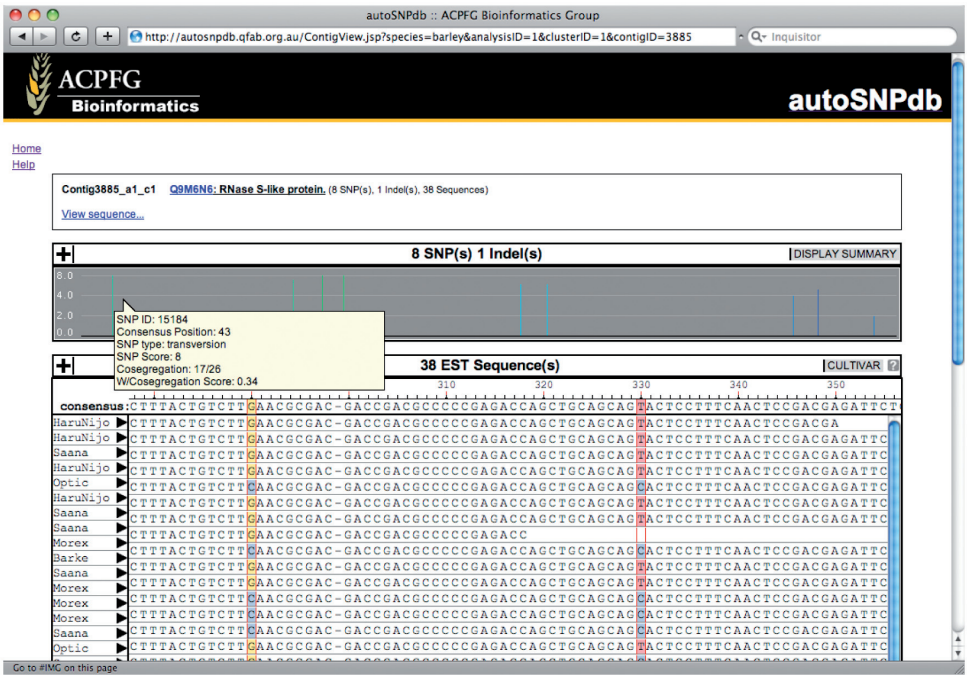

Figure 1. The autoSNPdb web interface displaying the sequence assembly, predicted SNPs as vertical bars and details presented in a mouse over box.

and novel queries specific to areas of interest. In the current version, we include plant cultivar information to allow the identification of SNPs that discriminate between plant cultivars.

METHODS

Data processing

Brassica, rice and barley expressed sequences were downloaded from GenBank release 159. RepeatMasker (www.repeatmasker.org) was used to identify and mask repeats prior to assembly using CAP3 (7) with the parameters –p 90, –o 50. The resulting assemblies and singleton sequences were parsed into a MySQL database. SNP discovery used the autoSNP method (4) implemented with custom Perl scripts, and the results were parsed to the database. Assemblies containing four sequences or more were examined for polymorphisms, with gaps created during the assembly process classified as indels. The minimum redundancy score defining a polymorphism was varied in proportion to the number of sequences in the assembly at the SNP position. A minimum redundancy score of 2 was required for up to seven sequences; 3 for between 8 and 11 sequences; 4 for between 12 and 19 sequences, and a minimum redundancy of 5 was required for predicting SNPs represented by 20 or more sequence reads. Each SNP was compared to all SNPs in an assembly to calculate the SNP co-segregation score, with the weighted co-segregation score calculated according to the proportion of missing data at that position in the assembly. Input sequences were annotated with cultivar type, tissue source and developmental stage where available. Consensus and singleton sequences were annotated based on sequence alignment using BLAST (8) against GenBank and UniRef90 databases. Gene Ontology (GO) annotations were derived from UniRef90 annotations. Comparative rice and Arabidopsis genome positions were derived by WU-BLAST comparison with TIGR rice pseudo-chromosomes (version 5) and TAIR Arabidopsis pseudo-chromosomes (v01222004), respectively.

Database content, access and interface

Barley, rice and *Brassica* sequences were downloaded from GenBank and processed through the autpSNPdb pipeline. A custom web interface allows users to query and visualize the SNP and annotation data (Figure 1). The maintenance of these data within a relational database enables numerous query options. Sequence annotations may be searched by gene keyword, sequence ID, GO term or through similarity to defined regions of the rice or Arabidopsis genome. A BLAST interface enables identification by sequence similarity. SNPs may be retrieved that differentiate between cultivars, providing a valuable resource for genetic mapping and association studies. To aid interpretation of the predicted SNP data, SNPs are viewed graphically as vertical bars, where the position of the bar along the *x*-axis reflects the relative position of the SNP in the consensus sequence; the height of the bar represents the SNP redundancy score; and the bar colour reflects the SNP-weighted co-segregation score. Information about each SNP is displayed by moving the cursor over the bar, while selecting a bar centres the sequence assembly at that position. The sequence assembly may be moved using the scroll bar and can be toggled between the full sequence assembly and a SNP summary. Labels to the left of the sequence may also be toggled between cultivars, GenBank accession numbers, tissue

type and development stage for the respective sequences. The interface is documented with help pages and database build information.

FUTURE DIRECTIONS

The autoSNPdb system was developed for flexible use and permits extension to a broad range of annotation and species. We plan to extend this system for other crops, including wheat and next-generation Roche 454 sequence data.

ACKNOWLEDGEMENTS

Support from The National Computing Infrastructure (NCI) and the Queensland Facility for Advanced Bioinformatics (QFAB) is gratefully acknowledged. This research was supported under Australian Research Council's *Linkage Projects* funding scheme.

FUNDING

Funding for open access charge: the Australian Research Council.

Conflict of interest statement. None declared.

REFERENCES

1. Syvanen,A.C. (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.*, **2**, 930–942.
2. 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.
3. 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. *Nat. Genet.*, **23**, 452–456.
4. 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.
5. 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.
6. Savage,D., Batley,J., Erwin,T., Logan,E., Love,C.G., Lim,G.A.C., Mongin,E., Barker,G., Spangenberg,G.C. and Edwards,D. (2005) SNPServer: a real-time SNP discovery tool. *Nucleic Acids Res.*, **33**, W493–W495.
7. Huang,X.Q. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
8. 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.