BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 19 2009, pages 2609–2610
doi:10.1093/bioinformatics/btp477

Sequence analysis

PineSAP—sequence alignment and SNP identification pipeline

Jill L. Wegrzyn[1,*], Jennifer M. Lee[2], John Liechty[1] and David B. Neale[1]

[1]Department of Plant Sciences and [2]Department of Evolution and Ecology, Address One Shields Ave. University of California, Davis, CA 95616, USA

Received on May 22, 2009; revised on July 31, 2009; accepted on August 2, 2009

Advance Access publication August 10, 2009

Associate Editor: John Quackenbush

ABSTRACT

Summary: The Pine Alignment and SNP Identification Pipeline (PineSAP) provides a high-throughput solution to single nucleotide polymorphism (SNP) prediction using multiple sequence alignments from re-sequencing data. This pipeline integrates a hybrid of customized scripting, existing utilities and machine learning in order to increase the speed and accuracy of SNP calls. The implementation of this pipeline results in significantly improved multiple sequence alignments and SNP identifications when compared with existing solutions. The use of machine learning in the SNP identifications extends the pipeline's application to any eukaryotic species where full genome sequence information is unavailable.

Availability: All code used for this pipeline is freely available at the Dendrome project website (http://dendrome.ucdavis.edu/adept2/resequencing.html)

Contact: jlwegrzyn@ucdavis.edu

1 INTRODUCTION

Single nucleotide polymorphism (SNP) detection involves looking across multiple sequence alignments and identifying base discrepancies. The higher the sequence coverage and quality score at a given site, the more confident the SNP prediction. In high-throughput studies, we rely on existing tools to make confident identifications as visual confirmation is not an option. Poor initial alignments can greatly increase both the false positive and false negative rate of SNP predictions. The automated alignment and SNP detection from re-sequencing data where a reference genome sequence is not available is an on-going challenge. Pine (Pinus) sequence data presents an even greater obstacle as it is highly polymorphic (average of one SNP per 50 bases) (Neale, 2007).

Existing programs such as Phrap (Lee and Vega, 2004) are heavily used for aligning genomic re-sequencing data and providing direct input to SNP prediction programs. Phrap, by definition, is a DNA sequence assembler and does not perform well when paired with the task of aligning highly polymorphic re-sequencing samples. It will often place individuals with different haplotypes into separate contigs. If the stringency of Phrap is reduced (in an attempt to force creation a single contig), misalignments of indels leads to poor overall sequence alignments. Quality benchmarks evaluated across several DNA and RNA aligners found ProbconsRNA (Do *et al*., 2005) to be highly accurate (Carroll *et al*., 2007; Wilm *et al*., 2006), however, it proves to be inhibitive in terms of speed for high-throughput studies.

SNP identification solutions that can accommodate fluorescence-based re-sequencing reads, often require genomic reference sequence. The utilities PolyPhred (Nickerson *et al*., 1997) and Polybayes (Marth *et al*., 1999) rely primarily on quality scores and sequence coverage. In tests utilizing Polybayes and Polyphred in loblolly pine (Pinus taeda L.), we received at best, 78% prediction accuracy with the majority of the discrepancy resulting from false positives. Recently, machine learning techniques have been applied to the problem of computational SNP discovery (Matukumalli *et al*., 2006; Unneberg *et al*., 2005; Zhang *et al*., 2005). Two of these applications rely on either an existing reference genome sequence (Zhang *et al*., 2005) or process EST sequences directly rather than raw tracefiles (Unneberg *et al*., 2005). The SNP-Phage application (Matukumalli *et al*., 2006) can call SNPs from fluorescent reads without a reference sequence, however, attention to the alignments of highly polymorphic organisms prior to SNP calling is not available.

PineSAP was developed as a high-throughput solution to analyze re-sequencing data in the form of chromatogram files for forward and reverse reads of multiple individuals. The pipeline presented here runs on the Unix/Linux platform and was written in Perl. PineSAP implements a combination of Phred, Phrap and ProbconsRNA to efficiently and accurately call bases and align re-sequencing reads. Following alignment, SNPs and indels are identified through the Polyphred and Polybayes packages. Sequence-based information is extracted and processed through a supervised machine learning algorithm for the purpose of accepting or rejecting the SNP predictions.

2 ALIGNMENT

The alignment section of the pipeline implements a de-coupled and modified version of phredPhrap on the original chromatogram files. Phred (Ewing *et al*., 1998) is responsible for the base calls and the assignment of quality scores. Phred is integrated into this pipeline with parameters to trim low quality ends in order to prevent alignment issues from bases in these regions. Conservative Phrap parameters are applied to prevent any misalignments in the resulting contigs and to ensure that all reads are retained.

Contig consensus sequences are exported from the ace format file generated from phredPhrap and aligned with ProbconsRNA. For each contig, an aligned FASTA file is created with each read in the contig aligned to the consensus sequence. These files along with the

*To whom correspondence should be addressed.

aligned file of all contig consensus sequences are used to generate a single multi-sequence FASTA file. In this file, each read is aligned to an overall consensus for the amplicon based on the alignment in the ProbconsRNA output and each read's alignment to the consensus sequence of its member contig. The final multi-sequence aligned FASTA file is converted back to an ace formatted file suitable for input to Polybayes and Polyphred.

3 SNP CALLING

The purpose of the classifier is to evaluate the accuracy of the SNP calls resulting from Polyphred and Polybayes. Sequence-based statistics were derived through a customized feature extraction program and fed as a vector for each polymorphism to the J48 classification tree available in the WEKA classifier package. The final set of features identified fully represents the local and global sequence variation, alignment depth and quality, local and global base quality and sequence alignment quality. From the set of 300 used for training, all of the true positive and false negative vectors are represented as real SNPs and the true negative and false positive vectors as non-SNPs. The resulting SNP calls were extracted from their respective output files flanking sequence, quality scores and a normalized confidence score.

4 VALIDATION

Manual validation of the alignments was completed with the same 300 loblolly pine amplicons. Each amplicon had between 8 and 36 reads. With Phrap alone and default parameters, 23% of the amplicons were placed in a single contig, 27% into two and 20% into three. Amplicons with two or more contigs could be forced into a single contig 82% of the time, however there are problems with the alignment in 34% of cases. The alignment method implemented in PineSAP improves the success rate to 98%.

When evaluated in terms of speed, the alignment method described above was run with a straight ProbconsRNA implementation. We determined it would take ∼25 times longer to process 36 sequences per amplicon.

The classification tree generated from the training sequences was tested against a unique set of 120 sequences with 563 manually validated SNPs. All SNP calls were identified as based on visual inspection of Polyphred and Polybayes predictions in Consed (Gordon *et al*., 1998). The classification tree resulted in a significant overall improvement with a calculated accuracy of 93.6% (Table 1).

Table 1. Results of SNP prediction on the test sequence data

Evaluation	J48	Polyphred	Polybayes
Accuracy	93.6	76.25	78.02
Sensitivity	88.21	83.22	86.54
Specificity	98.73	N/A	N/A

ACKNOWLEDGEMENTS

We would like to thank all the students that worked on the validation for both the alignments and SNP calls.

Funding: Association Genetics of Natural Genetic Variation and Complex Traits in Pine – NSF 0501763.

Conflict of Interest: none declared.

REFERENCES

Carroll,H. *et al*. (2007) DNA reference alignment benchmarks based on tertiary structure of encoded proteins. *Bioinformatics*, **23**, 2648–2649.

Do,C.B. *et al*. (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.

Ewing,B. *et al*. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Gordon,D. *et al*. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

Lee,W.H. and Vega,V.B. (2004) Heterogeneity detector: finding heterogeneous positions in Phred/Phrap assemblies. *Bioinformatics*, **20**, 2863–2864.

Marth,G.T. *et al*. (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genet.*, **23**, 452–456.

Matukumalli,L.K. *et al*. (2006) Application of machine learning in SNP discovery. *BMC Bioinformatics*, **7**, 4.

Neale,D.B. (2007) Genomics to tree breeding and forest health. *Curr. Opin. Genet. Dev.*, **17**, 539–544.

Nickerson,D.A. *et al*. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based re-sequencing. *Nucleic Acids Res.*, **25**, 2745–2751.

Unneberg, P. *et al*. (2005) SNP discovery using advanced algorithms and neural networks. *Bioinformatics*, **21**, 2528–2530.

Wilm,A. *et al*. (2006) An enhanced RNA alignment benchmark for sequence alignment programs. *Algorithms Mol. Biol.*, **1**, 19.

Zhang,J. *et al*. (2005) SNPdetector: a software tool for sensitive and accurate SNP detection. *PLoS Comput. Biol.*, **1**, e53.