PolySite: A tool for searching polymorphic sites within sequences from RNA-Seq and Sanger technology

Cassandra R[1], D'Agostino N[1], Traini A[1], Chiusano ML[1]

Motivation

The existing software for SNP detection work on few reads of short length (~ 100 nts) without providing a quality score and a clear classification of the detected polymorphic sites. We describe PolySite, a tool - written in Perl - that can effectively identify polymorphic sites within sequences from RNA-Seq as well as from Sanger technology. Compared to other software, PolySite is also able to discriminate polymorphisms due to sequencing errors. A key feature of this tool is that results are displayed into comprehensive, simple and intuitive html graphical output.

Methods

To assemble sequences into contigs, PolySite invokes cap3. For each contig, PolySite uses a "Minimum Redundancy Log Threshold" (LMRT) to assess the likelihood of a site to be polymorphic. The LMRT computation is based on log(n) where n is the number of sequences into each contig: in this way a site "is called" polymorphic according to the number of sequences in the cap3 multiple alignment. Then, based on LMRT, polymorphic sites are classified and likelihood values are associated to them using the "Quality Windows Method" (QWM). QWM is calculated considering all the assembled sequences into a single contig. If xi is the likely polymorphic site, the algorithm considers a window xi - # and xi + # and assigns a negative weight to any mismatch in the window. The QWM score represents a percentage describing the reliability of each polymorphic site. Using both LMRT and QWM, PolySite can also detect multiple polymorphic sites that could indicate heterozygosity or allelic multivariance as well as sequencing errors.

Results

Compared to other methods for SNP detection, Polysite is able to work on FASTA file including a large number of sequences (~ 550,000) obtained from next generation sequencing technologies or on ACE files generated by cap3. Many tests performed on different datasets from different species showed that polymorphic sites were correctly identified and classified and that possible sequencing errors were properly discriminated.

[1] Dept. of Soil, Plant, Environmental and Animal Production Sciences, University of Naples Federico II

Contact e-mail
cassandra.raffaele@gmail.com