Expressed Sequence Tags versus RNA-Seq. Methods and services for large-scale transcriptome analysis

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

Motivation

Since five years we have been involved in the analysis of expressed sequence collections mainly in the form of EST data and in the design of specific databases to support expression patterns detection from organism or tissue specific libraries. Recently, we expanded our methods to data from RNA-Seq which makes use of next-generation sequencing technologies. Here we discuss our strategies to enhance data quality and increase data information content as well as main novelties provided by our database collection which were designed thanks to collaborations within national and international projects with the aim to solve specific issues these data are useful for.

Methods

We implemented the ParPEST (Parallel Processing of ESTs) pipeline using public software integrated by in-house developed Perl scripts to process transcript data in the form of raw and partial transcripts. Input sequence data are screened i) for trimming low quality sequences ii) for cleaning contaminations and iii) for filtering and masking low complexity sub-sequences and interspersed repeats. Then, sequences are clustered and assembled to detect sequence redundancy and to generate gene indices. Automated annotation is obtained based on BLAST similarity searches against protein databases. For the description of sequence function, Gene Ontology (GO) terms and Enzyme Commission (EC) numbers are assigned to directly classify gene products according to international standards and to map on the fly the expressed sequences onto KEGG metabolic pathways. Similar methods have been adapted to manage larger collections as the one obtained from Next Generation sequencing technologies. Procedures were also introduced to better evaluated data quality and obtain from these sequence collections the most of their "deeper" sequence content.

Results

Transcript collections are certainly no substitute for a whole genome scaffold and show high levels of sequence redundancy and low quality sequence attributes. However, they currently represents the core foundation for understanding genome functionality and the most attractive route for broad sampling of transcriptome from

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

specific libraries. Thanks to the experience we gained in collecting, analysing and managing EST sequences, our interest naturally evolved with the evolving technologies. We obtained greater mastery in managing this type of data and we present our experience and results concerning the analysis of RNA-Seq data obtained from the Roche 454 platform, highlighting main challenges both in data management and analysis.

Contact e-mail
nunzio.dagostino@gmail.com