Sequence analysis of deep sequencing data of viroid-derived small RNAs in plants

Tulipano A[1], Navarro B[2], Di Serio F[2], Flores R[3] ,Gisel A[1]

Motivation

Viroids are circular, highly structured, non-protein-coding RNAs that are able to replicate and move through infected plants. Similarly to viruses, viroid infections are associated with the accumulation of viroid-derived 21–24 nt small RNAs (vd-sRNAs) with the typical features of the small interfering RNAs characteristic of RNA silencing, a sequence-specific mechanism involved in defense against invading nucleic acids and in regulation of gene expression in most eukaryotic organisms. To gain further insights on the genesis and possible role of vd-sRNAs in plant-viroid interaction, sRNAs isolated from different plants infected by host specific viroids were sequenced by the high-throughput platform Illumina. Algorithms to analyse these large data sets on such a specific task were not existent and we developed a semi-automatic analysis pipeline to analyse vd-sRNAs in plants.

Methods

Normally a deep-sequencing output contains the raw sequence data in FASTQ format, in the case of sRNA sequencing including the whole or a fraction of the 3'-adaptor sequence. In some cases the sequencing facility provides already a 'cleaned' output where the adaptor sequence was removed and the insert sequences were sorted by sequence size into different files. Starting with the raw sequences, the system extracts the sequences that contain a part of the 3' adaptor sequence and splits the results in different files according to the sequence size and provides corrected FASTQ and standard FASTA files of adaptor-free reads. FASTA output files contain unique sequences with their frequency noted in the header line. Sequencing with 36 cycles, the maximal insert size accepted from the system is 26 nt since at least 10 nt of the adaptor sequence is needed to identify it. The system provides in addition the distribution of the extracted sequence sizes, in our case with the highest numbers for the 20-, 21- and 24-mer sRNA. The extracted insert sequences are then mapped file by file (sequence sizes of interests) onto the corresponding viroid genome using the FASTA sequence files following two different strategies. If we deal with one genomic viroid sequence we use maq (Li et al, 2008) as a fast and reliable mapping tool, if we deal with several variants of the genome sequence, which is very often the case for viroids, we preferred to run a blastn (Altschul et al, 1997) against all genome sequences and

[1] Istituto Tecnologie Biomediche (ITB-CNR), Bari, Italy (2)Istituto di Virologia Vegetale, (IVV-CNR), Bari, Italy [3] Instituto de Biologà a Molecular y Celular de Plantas (UPV-CSIC), Valencia, Spain

extract from the blast output the best hits. The output file of this blast analysis has the same format than the maq output, so that both outputs can be processed further within the pipeline. In parallel the system splits the sequences that map onto the viroid genome(s) and the sequences that do not map and are sRNAs from the host plant. The vd-sRNA sequences are mapped for a graphical output onto the genome sequence(s) or onto the consensus of the genome sequences by mapping the frequencies of all insert sequences having the 5'-terminal nucleotide at a given position on the viroid genome. The mapping of these frequencies was done separately for (+) and (-) strand. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851-1858. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402.

Results

We presented a set of algorithms that specifically deals with the analysis of deep sequencing data from vd-sRNAs in plant systems, and that it can easily extended to analyse also virus-derived small RNAs. The system accepts raw sequences in FASTQ format as well as preprocessed sequence files in FASTQ format. Every intermediate step results in an output that is the input for the next step. The final result is a statistics about sequence size distribution of all isolated sRNAs, as well as only for the insert sequences mapped onto the viroid genome(s), and a mapping graphics showing the frequency of inserts initiating at a given position on the viroid genome. All intermediate files are kept for eventually further investigation. From the biological point of view we demonstrated that the large majority of vd-sRNAs derives from a few specific regions (hotspots) of the genomic (+) and (-) viroid RNAs. When grouped according to their sizes, vd-sRNAs always assumed a distribution with prominent 21-, 22- and 24-nt peaks, which, interestingly, mapped at the same hotspots. These findings show that different Dicer-like enzymes (DCLs) target viroid RNAs preferentially accessing to the same viroid domains.

Contact e-mail

andreas.gisel@ba.itb.cnr.it