An accurate pipeline for analysis of NGS data of small non-coding RNA

G. Giurato[1] ✉, M.R. De Filippo[2], C. Cantarella[1], G. Nassa[1], M. Ravo[1], E. Nola[3], A. Weisz[4]

[1]Laboratory of Molecular Medicine and Genomics, Faculty of Medicine and Surgery, University of Salerno, Italy
[2]Fondazone IRCSS SDN, Naples, Italy
[3]Department of General Pathology, Second University of Naples, Naples, Italy
[4]Division of Molecular Pathology and Medical Genomics, SS Giovanni di Dio e Ruggi d'Aragona Hospital, University of Salerno, Italy

Motivations

The discovery of various families of small non-coding RNAs (sncRNAs) in recent years revealed the complexity of the regulation of gene expression at both transcriptional and post-transcriptional level. Of the numerous sncRNAs, microRNAs (miRNAs) constitute a large family of 19-23 nucleotides long RNAs that participate in a variety of processes, such as cell development and differentiation, apoptosis and stress responses to carcinogenesis. Computational analysis indicates that a unique miRNA can regulate hundreds of genes, underlining the potential influence of miRNAs in almost every cellular pathway. Deep sequencing technologies provides a powerful strategy to explore miRNA populations (miRNA-Seq) with high sensitivity and specificity. Different computational approaches have been developed to analyze miRNA-Seq data, allowing to identify known and novel miRNAs, perform differential expression analysis and predict putative miRNAs targets. We combined these algorithms into an analysis pipeline and tested it on data obtained from our experiments in cancer cell lines.

Methods

The data obtained from the sequencer were filtered following several criteria. Since the sequence of the adapter is known, a Perl script was used to trim, from the raw data, the adaptors. The sequence reads were then filtered for quality and clustered to unique sequences to remove redundancy, retaining their individual read count information. Unique sequences 18 nucleotides or more in length were mapped, allowing up to one mismatch, on miRNA annotation according to miRBase version 18 using miRanalyzer. This detects the reads which correspond to known miRNAs, giving an estimation of expression level. miRBase repository is used because it offers information about mature (the mature sequence of known miRNAs), mature-star (the sequence which pairs with the mature miRNA in the miRNA secondary structure) and precursor miRNA sequences (sequence of the hairpin). miRNAs have been considered as expressed if they are detected at least 5 reads/sample. After detecting those that correspond to known miRNAs, the remaining reads were mapped to databases of transcribed sequences as mRNA and non-coding RNA (RFam). This step has two goals: (i) the number of matches can be viewed as a sample quality parameter (contamination of the RNA sample with degradation products and poly A tails) and (ii) it might be interesting to see which other known sncRNAs are in the sample. To predict novel miRNAs we used a probabilistic algorithm, miRDeep2, based on miRNA biogenesis model, to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. This tool aligns sequencing reads to potential hairpin structures in a manner consistent with Dicer processing and assigns log-odds scores to measure the probability that hairpins are true miRNA precursors. To detect novel miRNAs by miRDeep2, a score cutoff corresponding to a prediction signal-to-noise ratio >10 was used. Identification of differentially expressed miRNAs was performed with the Bioconductor DESeq package. Starting from the expression values, the first step was to minimize the effect of the systematic technical variations, and then a negative binomial distribution model was used to test differential expression in deep sequencing datasets. Only miRNAs with a p-values less or equal to 0.05 and fold-change less or equal to -1.5 and greater or equal to 1.5 were considered as differentially expressed. Given the critical roles of miRNAs in regulating gene expression and cellular functions, we predicted their putative targets, intersecting results obtained from two resources, TargetScan and microRNA.org. TargetScan provide computationally predicted miRNA gene targets by searching for the presence of 8 and 7 mer sites that match

the seed region of each miRNA, while microRNA.org target prediction incorporates current knowledge on target rules and on the use of a compendium of mammalian miRNAs. A further step of the analysis was to investigate nucleotide variations relative to the reference genome. To this purpose, preliminary steps were to reduce alignment artifacts and compute a more accurate quality estimation, removing biases due to sequencing cycle and preceding nucleotide. Further evidences were used to identify new miRNA variation sites: (i) Sequencing depth of variation sites should be equal to or larger than 5 reads per site, (ii) frequency of Single Nucleotide Variant occurrence >5% and (iii) variants not found in previous SNP annotation databases, like dbSNP.

Results

We developed an accurate pipeline for integral analysis of next generation sequencing data of small RNA molecules. Based on solid statistical methods, this allows both detection of known miRNAs and prediction of new miRNAs, integrating steps for differential analysis, sequence analysis and target prediction.

Acknowledgements

Research support by: Fondazione con il Sud; Italian Association for Cancer Research; Italian Ministry for Education, University and Research; Regione Campania; University of Salerno; Fondazione Veronesi. Giorgio Giurato is a student of PhD School in Experimental and Clinic Medicine / Doctorate in Experimental Physiopathology and Neuroscience, Second University of Naples. Maria Ravo is supported by a 'Vladimir Ashkenazy' fellowship of Italian Association for Cancer Research. Concita Cantarella and Giovanni Nassa are fellows of Fondazione con il Sud.