Non-coding RNA bioinformatics platform for full backing of the high-throughput sequencing experiments generated by Next-Generation Sequencing technologies

F. Licciulli✉, A. Consiglio, G. De Caro, A. Gisel, G. Grillo, A. Tulipano, S. Liuni
Istituto di Tecnologie Biomediche del Consiglio Nazionale delle Ricerche, Bari, Italy

Motivations

Short non-coding RNA molecules (20-30 nucleotides long) play an important role in the regulation of gene expression by interacting with their target RNAs. This interaction generally downregulates gene expression either affecting RNA stability or repressing translation. Different classes of small regulatory non coding RNAs (sncRNAs) have been discovered and studied so far, and new families continue to be described, which differ in the proteins required for their biogenesis, the mechanism of target recognition and regulation, and the biological pathways they control [1,2]. In particular, three major classes of sncRNAs have been mostly investigated: small interfering RNAs (siRNAs), micro-RNAs (miRNAs) and PIWI-interacting RNAs (piRNAs) [1,2,3]. siRNAs direct the endonucleolytic cleavage of their target RNAs through a mechanism known as RNA interference (RNAi), miRNAs can repress translation or direct degradation of their target mRNA generally through imperfect complementary pairing on their 3'UTRs, whereas the major role of piRNAs is to ensure germline stability by repressing transposable elements (TEs). Recently, the advent of new Next-Generation Sequencing (NGS) tech-

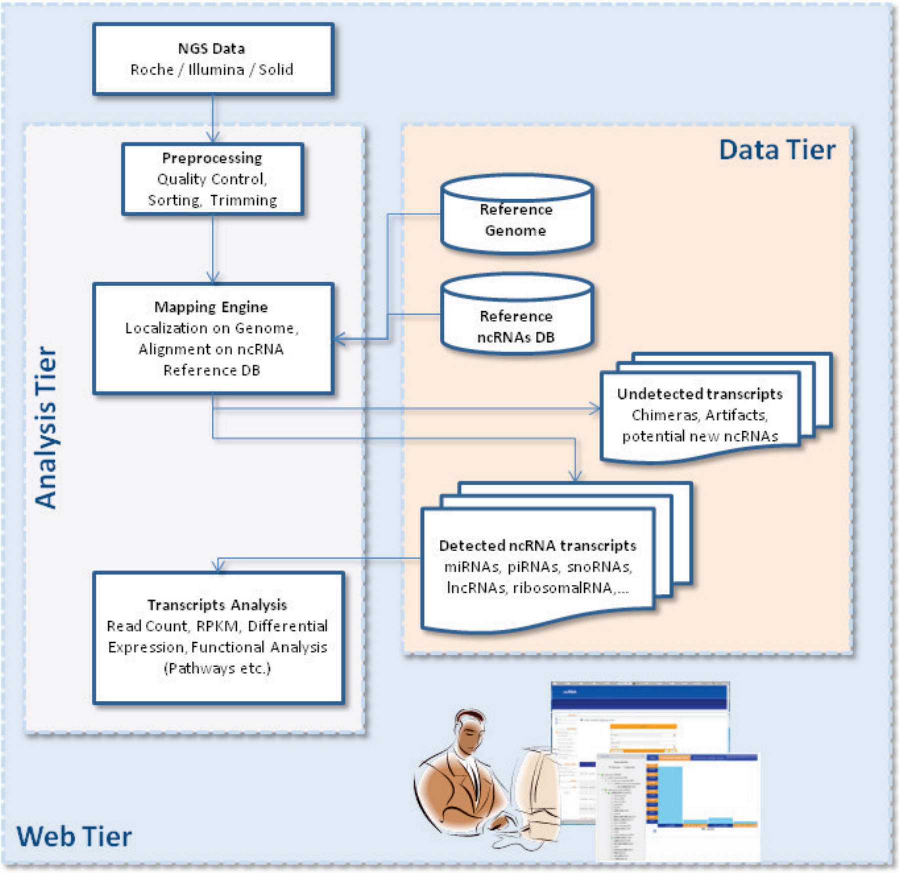

Figure 1: ncRNAs Platform Architectures

nologies has awfully increased the throughput of transcriptome studies, thus allowing an unprecedented investigation of non-coding RNAs. Regulatory pathways involving ncRNAs, such as miRNAs, are now being elucidated in detail and functions for long non-coding RNAs are also emerging. The huge amount of transcript data produced by high-throughput sequencing requires the development and implementation of suitable bioinformatics workflows for their analysis and interpretation. Here we describe here a bioinformatics resource to classify and analyze the non-coding RNA component of human transcriptome sequence data obtained by different NGS platforms (Roche 454, Illumina and Solid).

Methods

The ncRNAs bioinformatics platform is organized according to a typical three tier architecture: an analysis tier for ncRNAs detection, classification and functional analyses; a data tier made up of a data-warehouse used to store the analysis results, the ncRNAs reference database (a non-redundant collection of ncRNAs sequence retrieved from fRNAdb, RNAdb, mirBASE, NONCODE and others), the reference genome and other useful annotation database like HGNC nomenclature [4], Sequence Ontology (SO) [5] and Entrez Gene; a web tier module for querying the analysis results and the annotation stored in the ncRNAs reference database. The core of the platform is the analysis workflow. In figure 1 we show the pipeline for classification and functional annotations of non-coding RNAs (ncRNAs) fraction obtained through high-throughput sequencing (HTS) experiments using different NGS technologies. The input data for the bioinformatics platform can be either the reads data obtained by different NGS platforms (Roche 454, Illumina and Solid) or previously mapped reads stored in users' SAM/BAM files.

Results

The ncRNA bioinformatics platform - through a combination of an analyses pipeline, a data-warehouse and a user-friendly web interface - is able to:

i. detect and classify reads in known functional ncRNA categories using Sequence Ontology classification, HGNC nomenclature, gene names and miRNA accessions;
ii. extract reads collections belonging to a given category for further analysis;
iii. quantify ncRNA expression based on annotations derived from different reference ncRNA databases;
iv. generate some statistics of expressed ncRNAs, indicating the RPKM (reads per kilobase of RNA model per million mapped reads) value for each Sequence Ontology class;
v. detect differential expression of ncRNAs between two conditions (i.e. normal/pathological);
vi. create a collections of interesting clusters of reads mapped on the genome but not detected as known ncRNA;
vii. filter out reads mapping to ribosomal RNAs and mtDNA transcripts;
viii. create a collection of unmapped residual reads (chimeras, artifacts, and contaminations).

References

9. Ghildiyal, M. and Zamore, P. D. Small (2009) "Silencing RNAs: an expanding universe". Nature Rev. Genet. 10, 94-108.
10. Malone, C. D. and Hannon, G. J. (2009), "Small RNAs as guardians of the genome". Cell 136, 656-668
11. Kim, N. V., Han, J. '& Siomi M. C. (2009), " Biogenesis of small RNAs in animals". Nature Rev. Mol. Cell Biol. 10, 126-139
12. Wright M '& Bruford E (2011), "Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature". Human Genomics, VOL 5. NO 2. 90-98.
13. Eilbeck et al.(2005), "The Sequence Ontology: A tool for the unification of genome annotations". Genome Biology 6:R44