miRNA Sequence data analysis pipeline

Analysis steps/methods for miRNA Sequencing

1. Preprocessing and read quality assessment Sample specific reads preprocessing from multiplexed flow cells from raw base calls and quality scoring and De-multiplexing Read quality assessment and filtering
2. Alignment and expression calls
3. Normalization, Differential expression analysis
4. Functional profiling a.Target prediction, IsoMIR identification and novel miRNA prediction
5. Visualization

MiRNA sequencing pipeline in detail

To facilitate processing of the vast amount of data produced by Next Generation Sequencing, we developed a pipeline that performs preprocessing of reads for quality assessment quality filtering and adaptor trimming of raw reads, then performs sequence read alignment against reference genome assembly and miRNA quantification and lastly performs expression and functional analysis.

Preprocessing and read quality assessment

Quality control step generates quality check reports in HTML/PDF format. The reports highlights the per sequence quality scores, sequence base quality scores, sequence GC and Kmer content, duplicate sequences, overrepresented sequences and overall sequence length distribution.

For the preprocessing of raw base calls and sample de-multiplexing, pipeline uses open source CASAVA 1.8.2 toolkit by Illumina to generate text-based fastq format files with biological sequence and their quality scores. For the raw sequence reads quality assessment pipeline uses open source software FASTQC 0.11.2 and FASTAX tool kit 0.0.13.2. Then reads with low quality ends are filtered with quality cut of Phred quality score 20 and ligated adaptor sequence reads were trimmed with minimum over lap of 5 bases and allowed error rate of 0.1 using open source tool called cutadapt 1.2.1. At this step, filters are also applied to retain sequences of minimum length of 15 base pair and discard sequences below 15 base pair length.

Alignment and miRNA quantification

Using short read aligner Bowtie 1.0.0, sequence reads were aligned to hg19 human reference genome assembly and best hits to the genome are identified using bedtools 2.17 and small RNA species were annotated by mapping with reference gene feature annotation table from UCSC/ Ensembl database.

The alignment step generates alignment files and HTML files with mapping statistics for the percentage of reads mapped out of total number of reads against reference genome/miRbase. This also generates the count matrix of total read count for each annotated miRNA/gene for each sample in text file format

For the miRNA quantification/enrichment reads were first aligned to RFAM database for hg19 (homo sapiens) to filter and profile other small RNA species in the samples and then reads were mapped for small RNA annotation against reference mature/precursor hg19 small RNA sequence database miRBASE using bowtie aligner. Then per miRNA read count matrix for each samples were generated using open source tools PICARD 1.84, samtools 0.1.19, bedtools 2.17 and shell/perl/R scripts.

Normalization, Differential expression analysis

For the expression analysis step pipeline uses using several R packages for data normalization and variance stabilization such as DESeq, EdgeR, LIMMA VOOM, maSIGPro, BitSeq, Cummerbund, timecourse.

1. For tidying and cleaning data pipeline uses dplyr package:
2. For counts per million based RPKM calculation, normalization and exploratory analysis pipeline uses edgeR, DESeq, maSIGpro R packages.
3. For time series correlation with miRNA expression changes pipe uses maSigPro and edgeR to calculate fold changes across time points.

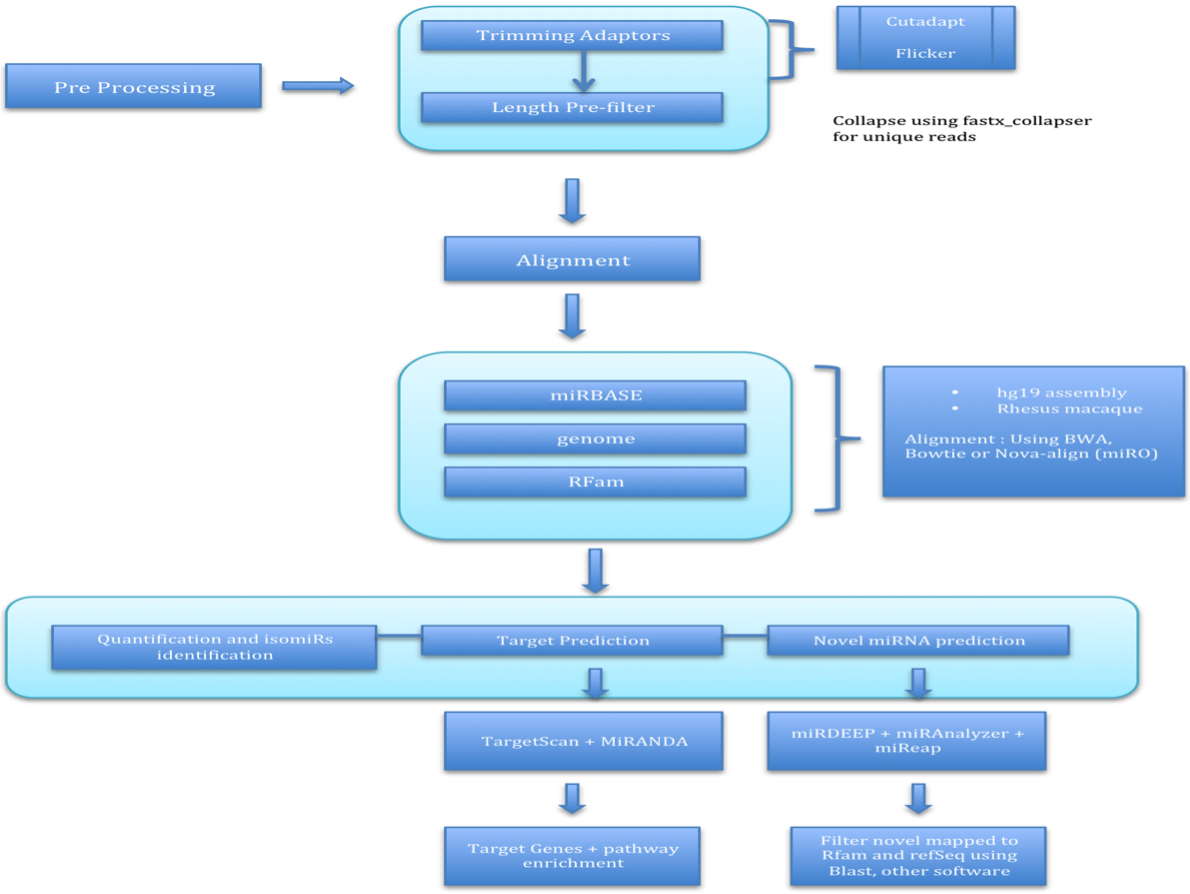
Functional profiling

TargetScan, miRanda algorithm, miRanalyzer and miReep tools will be used for target prediction, IsoMIR identification and novel miRNA prediction.

Visualization

For data visualization such as Heat Map, count profile bar plots for mapping statistics and RNA species bar plots, pipeline uses R/python functions and scripts.

miRNA seq Pipeline



miRNA seq Pipeline in detail

miRNA sequence analysis pipeline