MicroRNA analysis pipeline development for the next generation sequencing data

Simonas Juzėnas, Greta Varkalaitė, Ugnė Gyvytė, Laimutis Kučinskas, Jurgita Skiecevičienė, Limas Kupčinskas

Lietuvos sveikatos mokslų universitetas

MicroRNAs (MiRNAs) are a class of small non-coding RNAs involved in major carcinogenesis pathways. MiRNAs have been shown to exhibit potential diagnostic and prognostic properties in all major types of cancer. Next generation sequencing has become the main platform for MiRNA profiling. However, bioinformatic analysis of the sequencing data is challenging, as it requires significant amount of computational resources and currently available web based analytical tools lack flexibility and reliability. We developed an in-house sequencing pipeline for miRNA sequencing data analysis that integrates read pre-processing, alignment, mature/precursor/novel miRNA detection and quantification. Using well characterized data, we demonstrated the pipeline's superior performances, flexibility, and practical use in research and biomarker discovery.

MicroRNAs, RNA-seq, bioinformatics

Introduction

MicroRNAs (miRNAs) are widely studied small non-coding RNAs (~22bp) that regulate gene expression. Due to high stability of these molecules in biological samples, they have become an attractive target in the biomarker research field (Link and Goel 2013; Bartel 2009). Aberrant expression of miRNAs has been associated with a number of disease states, including cancer and autoimmune diseases (Link et al. 2012; Marques-Rocha et al. 2015; Kalla et al. 2015). Furthermore, miRNAs have been shown to have a diagnostic or prognostic role and even potential clinical implications for targeted gene therapy in cancer patients (Yamakuchi et al. 2010; Yanaihara et al. 2006).

MiRNA profiling through next generation sequencing (NGS) has become the main platform for biological research and biomarker discovery. However, analyzing miRNA sequencing data is challenging as it requires a significant amount of computational resources and bioinformatics expertise (Veneziano, Nigita, and Ferro 2015). Lack of flexibility and reliability of the currently available web based analytical tools is a common issue.

We aimed to develop an in-house pipeline for miRNA NGS data analysis that integrates read pre-processing, alignment, mature/precursor/novel miRNA detection and quantification.

Methods

Overview. The computational pipeline described in this paper represents an integrative toolkit for miRNA NGS data analysis. The overview of the methodology used in the pipeline is shown in Fig. 1. The bioinformatics pipeline starts by pre-processing the raw reads in FASTQ format: trimming 3' adapter sequences, quality filtering and collapsing identical reads to accelerate the computationally exhaustive downstream analysis. After the pre-processing step, the pipeline uses local alignment tool blastn to map the reads to a custom databases that contain annotated viral genome (O'Leary et al. 2016), viral hairpin (Kozomara and Griffiths-Jones 2014), tRNA, rRNA, snRNA and sRNA (Griffiths-Jones et al. 2003) sequences, which are then discarded from further analysis. The filtered reads then could be used for novel miRNA prediction or directly for known miRNA quantification using local alignment tool bowtie. Novel miRNA prediction step of the pipeline uses miRDeep2 (Friedländer et al. 2012) module. In this step filtered reads are mapped to the human genome (hg19) and used for novel miRNA prediction by miRDeep2 core algorithm. Predicted miRNAs are then filtered based on several criteria (signal to noise ratio; mapping to CDS, lncRNA and repeat sequences; GC content) and quantified as described previously. This pipeline is adapted to human miRNA NGS analysis but it can be used for both plant and animal miRNA analysis, including novel miRNA prediction, miRNA quantification and differential expression analysis.

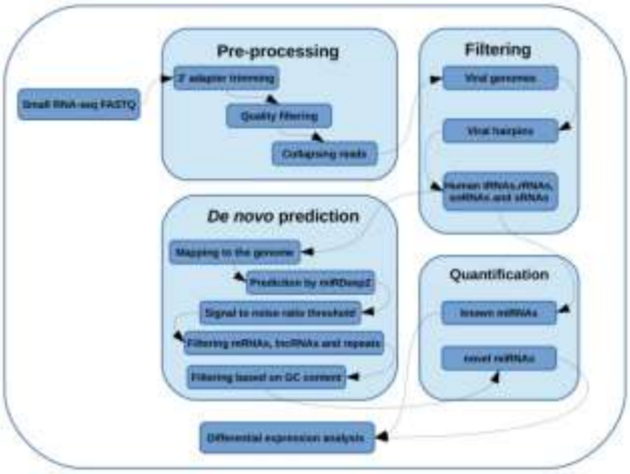

Fig. 1. *Scheme of computational pipeline for miRNA NGS analysis.*

Small RNA NGS data sets. In order to evaluate the performance of the computational pipeline, two different sources of material were used to generate small RNA libraries. 15 libraries were derived from paraffin embedded histologically normal human stomach tissue from patients with gastrointestinal stromal tumors and 24 libraries were derived from histologically normal freshly frozen human stomach tissue from patients with gastric cancer. All libraries were prepared using TruSeq Small RNA Sample Preparation Kit (Illumina) according to manufacturer's protocol and sequenced on HiSeq2500 (Illumina) next-generation sequencing platform.

Pre-processing sequencing reads. This step includes trimming 3' adapter sequences, quality filtering based on phred quality score (Q ≥ 20) and collapsing identical reads. Depending on the experimental design and used NGS platform, 3' adapter sequences may differ. The pre-processing step is adapted to TruSeq Small RNA protocol but it could be adjusted to other protocols. The 3' adapter trimming, quality trimming and collapsing is performed using the following code in the UNIX shell:

```
> for sample in `cat sample_list`; do
>      # Adaptor trimming and quality filtering
>      cutadapt -b TGGAATTCTCGGGTGCCAAGG -m 18 -q 20 --
discard-untrimmed ${sample}.fastq > ${sample}_clipped.fastq
>      # Fastq to fasta conversion
>      fastq_to_fasta -i ${sample}_clipped.fastq -o ${sample}_clipped.fa
>      # Collapsing identical reads
>      collapse_reads_md.pl ${sample}_clipped.fa >
${sample}_colapsed.fa
> done
```

Filtering sequencing reads. This step of the pipeline discards mapped sequences to annotated viral genomes, viral hairpins, tRNAs, rRNAs, snRNAs and sRNAs. The filtering is very important for *in silico* miRNA prediction in order to remove annotated sequences which could be false positively predicted as novel miRNAs. Even if novel prediction is not performed, it is recommended to filter the reads in order to detect possible contamination from other organisms. The example of filtering from viral genome sequences is shown below:

```
> # Create indexes of viral hairpins
> formatdb -i virus_hairpin.fa -pF -oT -n virus_hairpin.db
>
> for sample in `cat sample_list`; do
>      # Create indexes of reads
>      formatdb -i ${sample}_colapsed.fa -pF -oT -n
${sample}_colapsed.db
>      # Map sequences to viral hairpins
>      blastn -query ${sample}_colapsed.fa -db virus_hairpin.db -
word_size=18 -out  ${sample}_viral_hairpin.out
>     # Get sequence IDs that are not found in viral hairpins
>      get_noHits.pl ${sample}_viral_hairpin.out
${sample}_not_viral_hairpin_id.txt
>      # Retrieve not mapped sequences
>      fasta_cmd.py ${sample}_colapsed.db
${sample}_not_viral_hairpin_id.txt ${sample}_reads_filtered.fa
> done
```

Predicting novel miRNAs. The pipeline uses miRDeep2 module, developed to identify novel miRNAs from deep sequencing data. The algorithm uses a probabilistic-model-based method for miRNA discovery in animals. However, using modified parameters it was successfully applied for miRNA discovery in plants (Zhang et al. 2015). In the first step filtered sequences are aligned to reference genome by mapper module from miRDeep2 software which uses bowtie tool with the following options: bowtie –f –n 0 –e 80 –l 18 –a –m 5 –best –strata. These options allow 0 mismatches in the seed region of a read mapped to the genome (–n 0) and discard sequences which occur more than five times in the reference genome (–m 5). In the second step the RNAfold algorithm predicts RNA secondary structures of the potential precursors from reads which aligned to the reference genome. In the last step the potential novel miRNA precursors are scored or neglected by the miRDeep2 core algorithm (Friedländer et al. 2012).

The prediction step in the pipeline is performed using the following code:

```
> # Create indexes of human genome
> bowtie-build -f hg19.fasta hg19
>
> for sample in `cat sample_list`; do
>      # Map reads to reference genome
>      mapper.pl ${sample}_reads_filtered.fa -c -j -p hg19 -t
${sample}_reads_filtered_vs_genome.arf -o 2 -u -n
>      # Run prediction algorithm
>      miRDeep2.pl ${sample}_reads_filtered.fa hg19.fasta
${sample}_reads_collapsed_vs_genome.arf mirbase_mature.fa none
mirbase_hairpin.fa -t Human 2>report.log
> done
```

Quantification of known or/and novel miRNAs. The quantification of miRNAs is performed by quantifier module from miRDeep2 software. This module maps the sequencing reads to the known mature or novel miRNAs and their "star" sequences for the reference species against the known/novel precursor miRNAs for the reference species. The module uses bowtie with these options: bowtie –f –v 1 –a –best –strata –norc. The options allow 1 mismatch in the seed sequence (–v 1) and do not allow to map reads to the reverse complement of the precursor sequences in the reference (–norc). The quantification of known and novel miRNAs is performed using the following code in the UNIX shell:

```
> for sample in `cat sample_list`; do
>      # Quantify known miRNAs
>      quantifier.pl -p mirbase_hairpin.fa -m mirbase_mature.fa -r
${sample}_reads_filtered.fa -d
>      # Assign predicted mature and precursor sequences
>      dir=$(dirname ${sample}_reads_filtered.fa)
>      novel_hairpin=$(find $dir/mirna_results_*/ -type f | grep
"novel_pres.*fa")
>      novel_mature=$(find $dir/mirna_results_*/ -type f | grep
"novel_mature.*fa")
>      # Quantify novel miRNAs
>      quantifier.pl -p $novel_hairpin -m $novel_mature -r
${sample}_reads_filtered.fa -d
> done
```

Data and requirements for the pipeline:

- Operating system: Unix/Linux based.
- Programming language: Bash and Perl 5.
- Software: MiRDeep2, Cutadapt, FASTX, RNAfold, Bowtie, BioPerl and BLASTn.
- Databases: MiRBase, Rfam and Refseq.

Results and discussion

Pre-processing sequencing reads. To illustrate the performance of the pipeline, two different preservation types of histologically normal human stomach tissue were used for NGS miRNA analysis. It is known that fresh-frozen (FF) tissues usually have better RNA quality and yield than formalin-fixed, paraffin-embedded (FFPE) tissues (Roberts et al. 2009). Therefore, if the performance of the pipeline is good, this should be reflected in the results of pre-processing step. In total, small RNA NGS of 24 FF and 15 FFPE tissue samples yielded 198,993,444 (150,070,958 - FF and 48,922,486) raw sequencing reads. The average numbers of raw reads per sample were 7,898,471 and 3,261,499 in FF and FFPE samples, respectively. Whereas after the pre-processing, the average numbers of reads per sample were 6,879,945 and 2,180,589 in FF and FFPE samples, respectively. The average proportion of retained reads after pre-processing step was significantly higher (87,1 %; t-test FDR adjusted p-value = 0.00001) in FF than in FFPE (62,5 %) samples, which is as expected due to the lower quality of FFPE samples. The results of pre-processing show that all parameters of this step are well adjusted and sensitive enough to filter low quality reads from different RNA sources.

Filtering sequencing reads. Another step of the pipeline is to identify and discard various annotated sequences. This step is optional in the pipeline, however it is important for *in silico* novel miRNA prediction (Wen et al. 2012). After applying this step on the pre-processed reads, the average of filtered reads was higher in FF (2,253,447) than in FFPE (853,043) samples. However, the percentage of filtered reads was relatively similar in both groups of samples, 66,2 % and 60,1 % in FF and FFPE, respectively, showing that distribution of high quality reads is similar regardless the preservation type of biological material. The bulk of sequencing reads before and after filtering step were of 20-23 nt length which corresponds to the length of mature miRNA sequences. Interestingly, the majority of filtered reads were 26 - 43 nt length which exceeds the length of mature miRNA sequences (Fig. 2), showing that filtering step mainly discards uninformative sequences from downstream analysis. The greater number of these sequences were mapped to annotated transcripts from Rfam database, while only a small portion of sequences were mapped to viral sequences which could be used to identify viral infections in patients (Chang et al. 2013). The summary of filtered sequences is shown in Table 1.

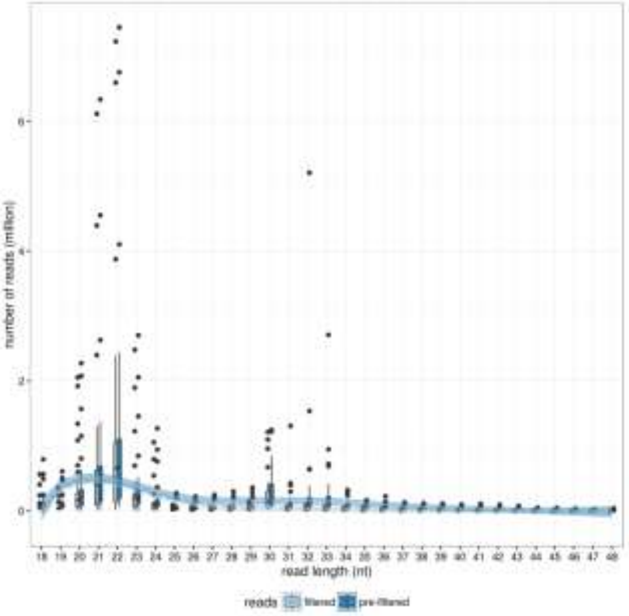

Fig. 2. *Pre-processed read lengths before and after filtering step.*

Novel miRNA prediction. The miRDeep2 is one of the most commonly used algorithm for miRNA prediction (Zhang et al. 2015; Wen et al. 2012; Shi et al. 2015). It identifies miRNA genes with high accuracy (98.6–99.9%) and sensitivity (71–90%) in all clades (Friedländer et al. 2012). Therefore it was chosen to implement in the pipeline. The input for *de novo* miRNA prediction is the sequencing reads after the filtering step. As an output, miRDeep2 produces a ranked list of novel candidates relying on the intrinsic features in terms of signal to noise ratio. In this step of analysis, 180 and 259 unique novel candidates above the signal to noise threshold value were identified in FF and FFPE samples, respectively. The average number of novel candidates per sample was 25 (ranged from 6 to 99) and 15 (ranged from 2 to 40) in FF and FFPE samples, respectively. Candidate novel miRNAs represented 153,773 read counts in total (range: 2 - 43,160), where 136,265 counts were detected in FF and 17,508 counts in FFPE samples. The predicted miRNA sequences were located in all 23 human chromosomes.

Quantification of known miRNAs. To demonstrate the the utility of the pipeline, known miRNAs from miRBase (version 21) were quantified in the sequencing reads retrieved directly after the filtering step. Overall, 1765 (1692 in FR; 1288 in FFPE) known miRNAs were detected and were represented by 103,662,405 (90,690,051 in FF; 12,972,354 in FFPE) counts. Interestingly, 73 known miRNAs were detected in FFPE were not detected in FR samples which had a higher yield of sequencing reads. However, the abundance of these miRNAs was very low (ranged from 0 to 4 counts per sample) and was not significant for downstream analysis.

In order to evaluate the overall performance of the pipeline, miRNA expression profiles of FF and FFPE samples from histologically normal stomach tissues were compared. Pearson's correlation analysis of miRNA mean expression values showed high similarity (r = 0.91) between FF and FFPE preserved tissues, meaning that the pipeline is sensitive enough to reconstruct similar expression profiles of the same origin of tissues even though the sampling and quality of RNA were different.

Table 1. *The summary of mapped sequences after filtering step*

Tissue preservation type	*The percentage of filtered sequences mapped to reference databases, %*		
	Viral hairpins	*Viral genomes*	*Rfam sequences*
Fresh Frozen (FF)	0.006	0.31	99.68
Formalin-Fixed, Paraffin-Embedded (FFPE)	0.003	0.63	99.32

Conclusions

In this report, we introduced a simple toolkit for miRNA NGS data analysis. The pipeline enables flexible and sensitive pre-processing, novel miRNA prediction and known/novel miRNA quantification from NGS data. Additionally, we demonstrated the pipeline's superior performances to reconstruct miRNA profiles from low quality RNA samples.

Funding

This work was supported by the Research Council of Lithuania under the initiative of Researcher teams' projects. Grant No: MIP-006/2014.

References

1. BARTEL, DP. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." Cell 136 (2): 215–33.
2. CHANG, ST., THOMAS MJ., SOVA P. et al. 2013. "Next-Generation Sequencing of Small RNAs from HIV-Infected Cells Identifies Phased Microrna Expression Patterns and Candidate Novel microRNAs Differentially Expressed upon Infection." mBio 4 (1): e00549–12.
3. FRIEDLÄNDER, MR., MACKOWIAK, SD., LI N. et al. 2012. "miRDeep2 Accurately Identifies Known and Hundreds of Novel microRNA Genes in Seven Animal Clades." Nucleic Acids Research 40 (1): 37–52.
4. GRIFFITHS-JONES, S., BATEMAN, A., MARSHALL, M. ET et al. 2003. "Rfam: An RNA Family Database." Nucleic Acids Research 31 (1): 439–41.
5. KALLA, R., VENTHAM, NT., KENNEDY, NA. et al. 2015. "MicroRNAs: New Players in IBD." Gut 64.
6. KOZOMARA, A., GRIFFITHS-JONES, S. 2014. "miRBase: Annotating High Confidence microRNAs Using Deep Sequencing Data." Nucleic Acids Research 42 (Database issue): D68–73.
7. LINK, A., GOEL, A. 2013. "MicroRNA in Gastrointestinal Cancer: A Step Closer to Reality." Advances in Clinical Chemistry 62.
8. LINK, A., KUPCINSKAS, J., WEX, T. et al. 2012. "Macro-Role of microRNA in Gastric Cancer." Digestive Diseases (Basel, Switzerland) 30 (3): 255–67.
9. MARQUES-ROCHA, JL., SAMBLAS, M., MILAGRO, FI. et al. 2015. "Noncoding RNAs, Cytokines, and Inflammation-Related Diseases." FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology 29 (9): 3595–3611.
10. O'LEARY, NA., WRIGHT, MV., BRISTER, JR. et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." Nucleic Acids Research 44.
11. ROBERTS, L., BOWERS, J., SENSINGER, K. et al. 2009. "Identification of Methods for Use of Formalin-Fixed, Paraffin-Embedded Tissue Samples in RNA Expression Profiling." Genomics 94 (5): 341–48.
12. SHI, J., DONG, M., LI, L., et al. 2015. "mirPRo-a Novel Standalone Program for Differential Expression and Variation Analysis of miRNAs." Scientific Reports 5 (January). Nature Publishing Group: 14617.
13. VENEZIANO, D., NIGITA, G., FERRO, A. 2015. "Computational Approaches for the Analysis of ncRNA through Deep Sequencing Techniques." Frontiers in Bioengineering and Biotechnology 3 (January): 77.
14. WEN, M., SHEN, Y., SHI, S., TANG, T.. 2012. "miREvo: An Integrative microRNA Evolutionary Analysis Platform for next-Generation Sequencing Experiments." BMC Bioinformatics 13 (1): 140.
15. YAMAKUCHI, M., LOTTERMAN, CD., BAO, C. et al. 2010. "P53-Induced microRNA-107 Inhibits HIF-1 and Tumor Angiogenesis." Proceedings of the National Academy of Sciences of the United States of America 107 (14): 6334–39.
16. YANAIHARA, N., CAPLEN, N., BOWMAN, E. et al. 2006. "Unique microRNA Molecular Profiles in Lung Cancer Diagnosis and Prognosis." Cancer Cell 9 (3): 189–98.
17. ZHANG, Z., JIANG, L., WANG, J. et al. 2015. "MTide: An Integrated Tool for the Identification of miRNA-Target Interaction in Plants." Bioinformatics (Oxford, England) 31 (2): 290–91.

Simonas Juzėnas, Greta Varkalaitė, Ugnė Gyvytė, Laimutis Kučinskas, Jurgita Skiecevičienė, Limas Kupčinskas

Automatizuotos algoritmų sekos sukūrimas mikroRNR sekoskaitos duomenų analizei

Santrauka

MikroRNR yra trumpos baltymo nekoduojančios RNR molekulės, dalyvaujančios su onkogeneze susijusių signalinių kelių reguliacijoje. Pasaulyje atlikti tyrimai parodė, kad mikroRNR raiškos profilis būna pakitęs įvairių vėžinių susirgimų atveju ir jos gali būti vertingais įrankiais šių susirgimų diagnostikai ir prognozei. Vienas patikimiausiu mikroRNR raiškos profilio tyrimo metodų yra naujos kartos sekoskaita. Tačiau bioinformatinė sekoskaitos duomenų analizė, reikalaujanti didelių kompiuterinių resursų ir patikimų bei lanksčių analitinių įrankių, vis dar išlieka iššūkiu. Šiame darbe pristatoma automatizuota algoritmų seka mikroRNR sekoskaitos duomenų analizei, kurioje yra integruotos pirminių nuskaitymų apdorojimo, seku palyginimo, subrendusių ir naujų mikroRNR bei jų prekursorių aptikimo ir kiekybinio įvertinimo funkcijos. Naudodami gerai charakterizuotus duomenis, pademonstravome sklandų sukurtos automatizuotos algoritmų sekos veikimą, lankstumą ir praktinį pritaikymą biožymenų paieškos tyrimams

MikroRNR, Naujos kartos sekoskaita, bioinformatika

Gauta 2016 m. kovo mėn., atiduota spaudai 2016 m. balandžio mėn.

Simonas JUZĖNAS. Lithuanian University of Health Sciences, Faculty of Medicine, Institute for Digestive Research, PhD student of Biomedical Sciences. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: (+370 37) 326092, e-mail: simonas.juzenas@fc.lsmuni.lt
Greta VARKALAITĖ. Lithuanian University of Health Sciences, Faculty of Medicine, Institute for Digestive Research, MSc student of Laboratory medical biology. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: (+370 37) 326092, e-mail: grevark@gmail.com
Ugnė GYVYTĖ. Lithuanian University of Health Sciences, Faculty of Medicine, Institute for Digestive Research, PhD student of Biomedical Sciences. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: (+370 37) 326092, e-mail: u.gyvyte@gmail.com
Laimutis KUČINSKAS. Lithuanian University of Health Sciences, Faculty of Medicine, Institute for Digestive Research, PhD of Biomedical Sciences. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: (+370 37) 326092, e-mail: laimisk@delfi.lt
Jurgita SKIECEVIČIENĖ. Lithuanian University of Health Sciences, Faculty of Medicine, Institute for Digestive Research, PhD of Biomedical Sciences. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: (+370 37) 326092, e-mail: jurgita.skieceviciene@lsmuni.lt
Limas KUPČINSKAS. Lithuanian University of Health Sciences, Institute for Digestive Research; Hospital of Lithuanian University of Health Sciences Kaunas Clinics, Department of Gastroenterology, MD, PhD, prof. Address: Mickevičiaus st. 9, LT-44307 Kaunas, Lithuania. Phone: Phone: (+370 37) 32 65 08, e-mail: limas.kupcinskas@lsmuni.lt