Last updated on 9/20/22 by RAA

MicroRNAseq processing pipeline

Trimmomatic v0.39 [1] and cutadapt v3.4 [2] were used in tandem to trim raw reads. Specifically, cutadapt was used to remove a commonly known miRNA specific adapter sequence, AGATCGGAAG [3]. Then, Trimmomatic was used to remove Illumina adapters and any reads with length less than 18 bp per usual miRdeep2 requirements [3]. FastQC v0.11.8 [4] and MultiQC 1.11 [5] were used to derive QC information and generate .html summaries both pre- and post- read trimming and alignment. The reference genome was indexed using BowTie [6] and BowTie2[7], and the mapper subscript of miRdeep2 (mapper.pl) was run with the following flags: -e -j -m -h. This enables the usage of FastQ files, autopurging of non-canonical bases, collapsing of reads, and parsing of Illumina FastQ to FastA respectively [3]. Collapsed reads were mapped using the miRdeep2 mapping script within miRDeep2.pl against MiRBase [3] and mature miRNAs. Resulting miRDeep2 predicted miRNAs and mature known miRNAs were used to map against with BowTie2 for quantification purposes. Samtools v1.3 [8] was used to filter alignments with quality less than 30 (-q 30) and to remove non-primary alignments (-F 2304) in order to generate per-sample miRNA raw counts. Proprietary BASH scripts were then used to combine all samples into a master raw miRNA counts table for the whole experiment while denoting known and miRDeep2-predicted putative miRNA nomenclature.

Required:

1) Raw single-end data in .fastq format
2) Reference genome and annotations

Deliverables:

1) QC reports
2) “Counts table” of raw miRNA counts per sample (known + miRDeep2 predicted)
3) miRDeep2 generated novel + known miRNAs and scoring table .html
4) Optional differential expression analysis via DESeq2

References:

1. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.
2. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. Available at: <https://journal.embnet.org/index.php/embnetjournal/article/view/200>. Date accessed: 20 sep. 2022. doi:https://doi.org/10.14806/ej.17.1.200.

3. Marc R. Friedländer, Sebastian D. Mackowiak, Na Li, Wei Chen, Nikolaus Rajewsky, miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades, *Nucleic Acids Research*, Volume 40, Issue 1, 1 January 2012, Pages 37–52, https://doi.org/10.1093/nar/gkr688.
4. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
5. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048 (2016).
6. Langmead, B., Trapnell, C., Pop, M. *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009). https://doi.org/10.1186/gb-2009-10-3-r25.
7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.
8. Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, Volume 25, Issue 16, 15 August 2009, Pages 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.

Note: For sequencing data acquisition please contact Emory Integrated Genomics Core (EIGC@emory.edu).

Questions? Comments?

Please email us at EICC@emory.edu